Part-based Grouping and Recognition: A Model-Guided Approach

Maurizio Pilu

Ph.D.
University of Edinburgh
1996

Abstract

The recovery of generic solid parts is a fundamental step towards the realization of general-purpose vision systems. This thesis investigates issues in grouping, segmentation and recognition of parts from two-dimensional edge images.

A new paradigm of *part-based grouping* of features is introduced that bridges the classical grouping and model-based approaches with the purpose of directly recovering parts from real images, and part-like models are used that both yield low theoretical complexity and reliably recover part-plausible groups of features. The part-like models used are statistical point distribution models, whose training set is built using random deformable superellipse.

The computational approach that is proposed to perform model-guided part-based grouping consists of four distinct stages.
In the first stage, *codons*, contour portions of similar curvature, are extracted from the raw edge image. They are considered to be indivisible image features because they have the desirable property of belonging either to single parts or joints.
In the second stage, small seed groups (currently pairs, but further extension are proposed) of codons are found that give enough structural information for part hypotheses to be created. The third stage consists in initialising and pre-shaping the models to all the seed groups and then performing a full fitting to a large neighbourhood of the pre-shaped model. The concept of pre-shaping to a few significant features is a relatively new concept in deformable model fitting that has helped to dramatically increase robustness. The initialisations of the part models to the seed groups is performed by the first direct least-square ellipse fitting algorithm, which has been jointly discovered during this research; a full theoretical proof of the method is provided.
The last stage pertains to the global filtering of all the hypotheses generated by the previous stages according to the Minimum Description Length criterion: the small number of grouping hypotheses that survive this filtering stage are the most economical representation of the image in terms of the part-like models. The filtering is performed by the maximisation of a boolean quadratic function by a genetic algorithm, which has resulted in the best trade-off between speed and robustness.

Finally, images of parts can have a pronounced 3D structure, with ends or sides clearly visible. In order to recover this important information, the part-based grouping method is extended by employing *parametrically deformable aspects models* which, starting from the initial position provided by the previous stages, are fitted to the raw image by simulated annealing. These models are inspired by deformable superquadrics but are built by geometric construction, which render them two order of magnitudes faster to generate than in previous works.

A large number of experiments is provided that validate the approach and, since several new issues have been opened by it, some future work is proposed.

Acknowledgements

Firstly, I wish to sincerely thank my supervisor Bob Fisher for his support and invaluable comments and patience, which have made me mature professionally. I gratefully acknowledge the partial support of SGS-THOMSON Microelectronics, which has been made possible by Ing. Eugenio Cavallari and Ing. Roberto Fantechi.

A. Fitzgibbon has been a continuous source of stimulating ideas; his help, both scientific and technical, has been truly priceless. I thank also D. Borges, with whom I had discussions on part-based recognition at very early stages of my PhD.

More specifically, I acknowledge A. Fitzgibbon for discussions on Section 3.4 and for convincing me of the significance of the idea, and for suggesting the use (and providing an in-house implementation) of a genetic algorithm for the optimisation phase of section 5.3.5; David Eggert for introducing me into the aspect-based philosophy of Chapter 6 and contributing to it with some perceptive comments; Peter Oliver and Simon Perkins, for doing part of the laborious proof-reading of my English; my supervisor, for spotting some flaws and, above all, lack of clarity in the early theoretical formulation of Section 5.3.3; Dr. Mehmet Firat, for suggesting to me to use Adaptive Simulated Annealing and providing a pointer to a good implementation; and, finally, all the test subjects that participated to the psychological experiment in Appendix E. Thanks also to all the other members of the vision lab, in particular A. Ashbrook, E. Bispo, P. Fillatreau, A. Gionis, A. Lorusso, B. Southall, D. Wren, and M. Wright.

Finally, and most importantly, the moral support of my beloved parents, my brother Roberto and my sister Maria gave me the strength and confidence to complete this work. Thanks.

Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research.

Maurizio Pilu
Edinburgh
September 9, 1996

Contents

List of Figures

Chapter 1

Introduction

Vision in one of the most amazing results of evolution. A high portion of the human brain has evolved in such a way that a startling amount of visual stimuli are processed in parallel and holistic information is extracted in an apparently effortless fashion.

To the layperson, vision is as easy as opening his eyes on a sunny morning. To the biologist, vision is as wonderful and inexplicable as the origin of life. To the computer vision scientist it is a great challenge, which often turns to frustration.

It is enough to hang around at any vision conference to discover how this happens: we are in a situation where specialists cannot even understand each other, so much are we lost in the mighty avalanche of approaches, techniques, frameworks and hyper-technical details constantly streaming out from the literature.

This condition is the result of the deadly cocktail of the disarming naivete of the discipline, the formidable difficulty of the problem and, last but not least, the delirious ease with which ideas, either bad or good, are spread nowadays.

Originality is therefore a rare flower in such a frantic state-of-the-art as much as usefulness is. New ideas often blossom from simple but profound, pioneering seeds. One of these seeds was planted several years ago: *part-based recognition*.

1.1 The problem investigated

The segmentation of objects into their constituent parts is, like its closely related problem of figure-ground segmentation, one of the hardest problems in vision. The final aim would be to segment objects from their background and decompose objects into parts, however complicated the background and the object might be. For performing such a fundamental task, humans employ all sorts of information such as edges, colour, shading, and prior and contextual knowledge.

This thesis deals with some issues related to how generic part models can be used to guide grouping and segmentation of object parts from two-dimensional edge images. The task that I set for this work was much humbler than that of all-purpose part segmentation: here, only edge information is used and the domain of parts dealt with is very simple, although with a fairly descriptive power. Figure 1.1 gives an example of the typical input – an unsegmented edge image – and the desired output.

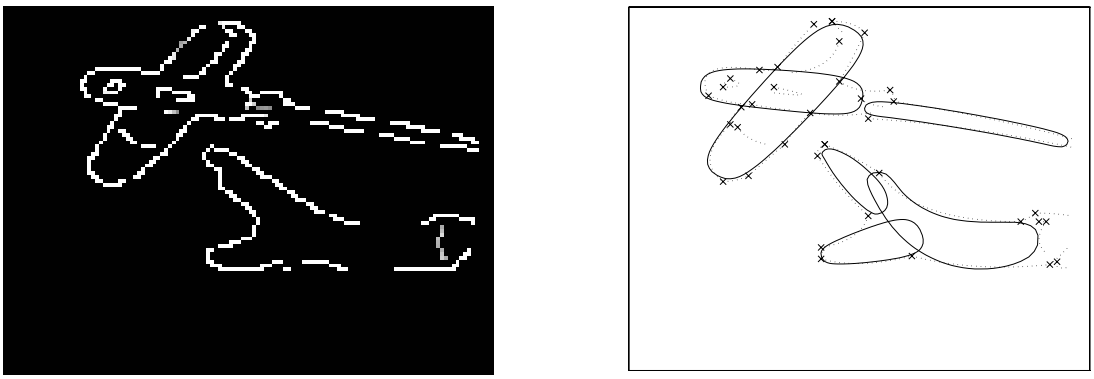


Figure 1.1: Left: The edge image input. Right: A typical part segmentation output.

The *leitmotiv* of the thesis is that for simple but relatively general cases like these, *generic models can be used to guide all processing stages, from locating and grouping features to segmentation and recognition.*

The purpose of the research was not to develop an improbable generic part recognition system but to investigate and propose solutions to some of the key issues.

1.2 Issues and proposed solutions

The task of performing segmentation from two-dimensional images by generic models is a challenging one and several issues have naturally arisen throughout.

In the following I introduce the main issues and outline the solutions that have been investigated in this thesis.

1.2.1 Modelling generic parts

In the proposed model-guided part recognition philosophy, the projected contour of parts need to be represented by models. The choice of the kind of model is a critical one, because it affects the recovery strategy: some models are easier than others to recover from real imagery. Parametric models have been widely accepted as a convenient way of representing part shape in a compact and natural fashion. In Section 3.4, I propose the use of a parametric model that is inspired by deformable superellipses (Sec. 3.3), but offers several advantages over them, notably its linearity and ease of fitting. Other related issues have been investigated, in particular a major contribution regards the first ellipse-specific direct least squares fitting algorithm of Section 3.2.

1.2.2 From edges to part hypotheses

Having immediately ruled out the possibility of producing direct part segmentation, for it would have required too strong assumptions on the quality (and probably content) of the input edge image, the strategy that has been followed is to first find a number of hypotheses and then filter them to produce a globally good segmentation into parts. Many new interesting issues have come up and a method is proposed in Chapter 4.

As is well known from the literature, the fitting of deformable models is either done, with very few exceptions, by manual initialisation to unsegmented data or to already segmented data. In this thesis the undertaking was to propose a way to overcome this limitation, at least in the domain of simple part models.

The fundamental idea (4.3) is to use small groups of *codons*, pieces of contour having similar curvature, that have the property of belonging to single parts, as *seeds*

from which the generic deformable model could be first pre-shaped and then fitted to additional evidence found in the image. This approach of choosing a small set of representative features, a familiar one in the traditional vision literature, has seldom be used for deformable models. The phase of choosing small groups of codons heavily affects the overall computational complexity of the method. I have used pairs of codons for this purpose, but future work is proposed that would use convex grouping as a better method to generate such small sets.

1.2.3 From part hypotheses to part segmentation

Once part hypotheses are available, there is the following problem: Is it possible to filter sets of hypotheses in order to yield the ones that are most likely to correspond to actual object parts?

A simple-minded approach for tackling this problem would be to retain hypotheses that have enough salience but, as shown in Section 5.2, this approach does not fulfill the simple requirement of producing a globally minimal representation.

A significant method, presented in Section 5.3, has been studied and implemented that tries to globally account for contending hypotheses under the Minimum Description Length paradigm. This method can be seen as producing the best interpretation of the edge data in the "language" of the generic part-models. The results are very encouraging but some principled limitations have been discovered, of which full account is given.

1.2.4 Recovering coarse 3D structure

The proposed method yields the segmentation of an image into 2D parts or outer contour of 3D parts. However, images of parts can have a pronounced 3D structure, with ends or sides clearly visible and if we want to recover this qualitative information, a true 3D model has to be employed. The fitting of three-dimensional generic models to 2D unsegmented images is a rather unexplored topic and no clear solution was readily available when this necessity come up.

This issue has been investigated and the proposed solution (Chapter 6) has been to

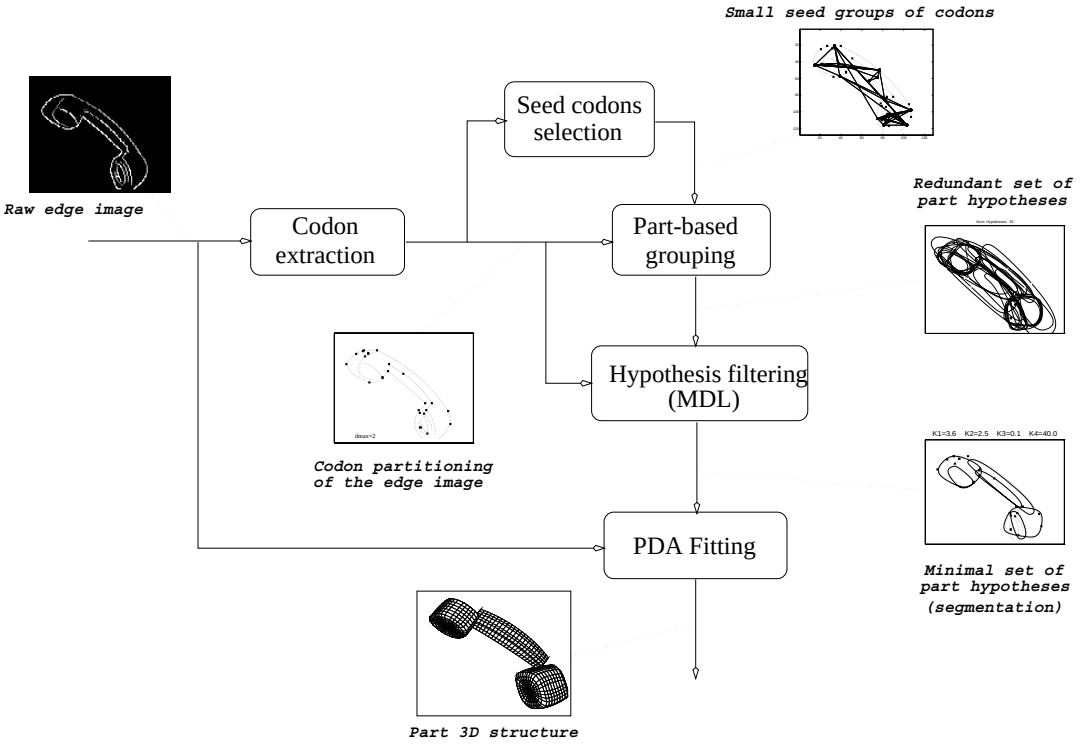


Figure 1.2: The structure of the conjectural part-based vision system that would use the stages described in this thesis.

employ *Parametrically Deformable Aspects* (PDA) models which, starting from the initial position provided by the previous stages, are fitted to the raw edge image by a theoretically sound Maximum a Posteriori estimation of a cost function inspired by information-theoretical arguments.

The PDA models resemble deformable superquadrics but are efficiently built by geometric construction which directly render their projected contour.

1.3 An hypothetical vision system

As said before, this thesis investigates issues concerning the model-driven strategy for part grouping and segmentation from 2D images, and does not describe a whole part-recognition system. However, the problems investigated and the proposed solutions were inspired by keeping in mind a possible part-based vision system. This not only has given a clear direction to the thesis, but also shaped its multi-stage structure.

Figure 1.2 depicts the structure of this hypothetical system and at the same time shows how the different topics discussed in this thesis relate to each other. From the raw input edge image, codons are extracted and then used to form small seed groups that will allow generic part models (the generic part Point Distribution Model of Section 3.4) to be initialised (by the ellipse-fitting of Section 3.2) and then fitted to additional codon evidence. Many hypotheses are produced by this grouping stage (Chapter 4), which are subsequently reduced by the Minimum Description Length part filtering stage (Section 5.3). Once part segmentation is available, qualitative 3D structure is recovered by the final parametrically deformable aspect fitting stage (Chapter 6).

1.4 Structure of the thesis

The thesis is structured as follows.

Chapter 2 gives a general background on part-based recognition. Section 2.1 provides a review of the concept of part as it is seen in cognitive psychology and Section 2.2 concentrates on the origin and developments of the part-based recognition idea in computer vision. Section 2.3 describes possibly the most widely recognised theory of part-based recognition, which is called Recognition by Components. Finally, Section 2.4 presents a rather detailed description and discussions on the three most popular approaches to part-based recognition to date.

Chapter 3 discusses modelling and model fitting issues. Section 3.2 presents a theoretical account of the first ellipse-specific direct least squares fitting method. Section 3.3 presents the deformable superellipse model, used here only for training a Point Distribution Model, which is presented in Section 3.4.

Chapter 4 addresses the problem of producing part hypotheses out of an unsegmented edge image. Section 4.1 and Section 4.2 give more specific (than the one of Section 2.1) background and literature review related to the proposed approach, which is outlined in Section 4.3. The following Sections 4.4, 4.5 and 4.6 detail the approach and 4.7 provides abundant experimental evidence. Contributions and criticisms to the method are discussed in Section 4.8.

Chapter 5 discusses the proposed hypothesis filtering method. After the illustration in Section 5.2 of a simple method for sorting hypotheses by their perceptual saliency, an interesting approach based on the Minimum Description Length criterion is fully detailed in Section 5.3. The chapter concludes with a example on how more information could be integrated for solving certain kind of ambiguities in Section 5.4 and with a discussion on the contributions, limitations and future work in Section 5.5. Remarkably, in Appendix E this segmentation results are compared to those obtained from a purposely-devised psychological experiment in which a number of voluntary subjects were asked to segment out parts from the same test images.

Chapter 6 deals with the problem of recovering the tri-dimensional structure of parts by fitting Parametrically Deformable Aspects (PDA). After a review of related work in Section 6.2, the building of the PDA is detailed in Sections 6.3 and 6.4 and the fitting method in Section 6.5. The experimental set-up and some results are shown in Sections 6.6 and 6.7, respectively. The chapter comes to a conclusion with a discussion of the proposed method.

Finally Chapter 7 concludes the thesis with an overview of the contributions, criticisms and some proposed future work.

Chapter 2

Background and Previous Works

This chapter gives a general background on part-based recognition. First, I review the concept of part from a cognitive psychology perspective and next its origin and developments in the computer vision field. Section 2.3 will extensively review and comment on a relatively recent part recognition theory, due to Biederman. Finally, Section 2.4 reviews some relevant previous work in part recognition.

2.1 The use of categories and parts in human vision

Perception is our window on the world: it allows us to recognise objects, relations, situations in the surrounding environment. In the case of vision, for doing this the eye-brain system must be able to properly transform, analyse and interpret a two-dimensional array of light intensity data. We are able to perform rapid recognition from a single image of scenes with many, possibly unknown or partially occluded, objects from a wide range of viewpoints and with different lighting conditions. This formidably complex behaviour has been investigated for many years by psychologists but we are still very far from any satisfactory explanation of the phenomenon.

It seems rather clear, however, that we exploit environmental regularities [Wertheimer 23, Gibson 79, Lowe 85, Pentland 86], like rigidity, objectness, etc. As Pentland [Pentland 86] reported from [Rosch 73b], *"if the apparent complexity of our environment were approximately the same as its intrinsic (Kolmogorov) complexity, then intelligent prediction and planning would be impossible, for there would not be*

any lawful relations. It is this internal structuring of our environment, then, that causes object features to cluster into groups, and allow us to successfully reason using the simplified category description we typically employ". This basic capability of our visual system to derive relevant grouping and categories without *a priori* knowledge of the content of a scene is called *perceptual organisation* [Lowe 85]. We could not perceive and recognise if we could not pick out the *essential* and separate it from the *accidental* [Bajcsy & Solina 87]: this is what is called *categorisation* by psychologists.

There are two main theories of categorisation that differ according to their membership rule. The first claims that membership of a category is determined by satisfaction of a set of properties or features (e.g. [Harnad 87]) while the other says that categories contain one member that is most representative and other members are perceived in terms of the prototype (e.g. deviations). The prototype theory is mainly due to Eleanor Rosch [Rosch 73a] and was soon picked up by artificial intelligence scientists such as in [Minsky 75] and [Marr & Nishihara 78]; now this theory is the one commonly adopted by the computer vision community. However, later (e.g. in [Rosch 78]) Rosch moved towards a less radical and simplified view: *"[...] the representation evoked by the category was more like good examples than poor examples of the category. [...] researchers have considered the concept of prototypes and typicality functions under-specified and have provided a variety of precise models, mini-models, and distinctions to be tested."* [Lakoff 87][1].

Psychological experiments in support to the prototype theory show that prototypes are more rapidly recognised than other objects [Rosch 73a, Biederman 87]. On the other hand, the feature theory lacks representational power, since by using context dependent features (e.g. an animal has four legs and a tree one) it is possible to obtain a better and better separation into categories but no "essence" [Wittgenstein 53] of what we are categorising would be captured.

[1] Further discussion on this matters are beyond the scope of this thesis. Refer to [Lakoff 87] for a excellent treatise on categorisation theory and its fascinating aspects.

There are three types of categories [Rosch 73a]:

Basic categories, closely linked to the structure of the perceived world.

Superordinate categories, whose members are basic categories that seem to share prevalently functional features.

Subordinate categories, which subdivide basic categories according to a few perceptual or functional features.

Examples of these categories are a chair, a piece of furniture, and an armchair, respectively.

Basic categories seems to have a privileged role in our perception system; it has been shown that children learn them first and they are recognised faster than super and subordinate categories (see, e.g., [Rosch 73b], but everybody's experience could also confirm that).

The hypothesis of the use of parts in human vision could be seen as a consequence of the prototype theory of categorisation, as *"it seems that the right level of* granularity *for representing basic categories seems to be primitives that correspond to the human notion of* part*"* [Bajcsy & Solina 87].

The human notion of "part" is somewhat fuzzy, as most people cannot give a coherent definition by introspection. However, we have carried out a psychological experiment in order to understand whether people do at least agree on what parts are, given a certain image. We provided as input just edge images[2] and the test subject had to decide about the decomposition into parts of the objects therein. The results are very supportive, at least to the claim that the notion of part has a neat role in the cognitive process. The full description of the experiment and its results are given in Appendix E.

[2] The other aim of the experiment was to compare their judgement to the part decomposition produced by the method that is going to be described in this thesis, which uses edge images as input.

2.2 Use of parts in computer vision

In the search for the building blocks of the visual perception system, the concept of basic categories and prototypes in the spirit of [Rosch 73a] became of prominent interest to computer vision researchers, because they would give perceptual salience and biological plausibility to possible modelling primitives (which were both missing in the old CAD-like paradigms) towards more general vision systems. As Lowe pointed out, *"why should we be constrained by the biological solution to the problem? [..] Because biological vision is currently the only indication we have that the general vision problem is even open to a solution"* [Lowe 85].

Primitives, seen as basic categories, offer a very general representation paradigm for computer vision [Biederman 87, Pentland 86, Marr & Nishihara 78] and, moreover, as they are *non-overlapping*[3] they could also do very well as model data base; super and sub-categories can be built starting from the basic categories by grouping or by more detailed analysis. This view, as noticed in [Pentland 86], is somehow opposite (it is more top-down) to Marr's scheme of successively describing images, edges, surfaces and volumes [Marr 82].

It also seems that the concept *we* have of *part* is well suited as a primitive for the construction of basic categories of objects [Pentland 86, Bajcsy & Solina 87, Lowe 85, Hoffman & Richards 85, Marr & Nishihara 78].

The key problem of this strand of research was to find a set of part models that comply with requirements of generality and detectability, and *the search for image regularities that are lawfully associated to these parts* [Lowe 85]; the content of the image would then be expressed as a combination of these primitives. In this respect, the representational power of these parts will be less than that of the whole object but greater than that of surfaces, contours or points; moreover they must be *"complex enough to be reliably recognisable, and yet simple enough to be used as building block for specific objects"* [Pentland 86].

Several part models have been proposed or used in the past. In the '1960s

[3] This term was used in [Biederman 87] to indicate primitives which are defined by the Cartesian product of qualitative properties.

and early '70s, the concept of objects made up by simple polyhedral parts was introduced in [Roberts 65] and successively further investigated and refined in [Clowes 65],[Huffman 71],[Waltz 75] and others. These methods have had limited success in industrial and some other applications where there is prior knowledge of the environment and the objects therein.

Binford introduced the idea of generalised cylinders [Binford 71] and later Brooks implemented a system in which generalised cones were used as generic modelling primitives [Brooks 84]; however, except from few examples, they turned out to be too general to be easily detected from an image.

An improvement towards more detectable primitives was suggested in [Pentland 86] by using superquadrics (a subset of generalised cones) and later in [Solina & Bajcsy 90] and in [Raja & Jain 92b] with deformable superquadrics; however, most of the research with these primitives is concerned with recognition from range data

Lowe was one of the first computer vision scientists that tried to investigate our perceptual organisation in order to find features and relations to be used for detecting significant components of objects [Lowe 85]. His work was very successfully followed up by Biederman, who developed his theory of Recognition by Components (see next section) and deliberately proposed a specific set of parts called *geon*s that arises from the exploitation of non-accidentalness in two-dimensional images [Biederman 87].

The concept of basic categories and parts is not the panacea for solving either computer vision or the understanding of human vision; rather, it should be considered as a significant contribution to our understanding of the basic principles of vision. From an engineering standpoint, a *wise* choice of the kind of parts in a part-based recognition framework could dramatically reduce the difficulties currently faced in dealing with real images.

2.3 Recognition by Components

The paradigm of Recognition by Components (RBC), introduced by [Biederman 87], provides a link between studies on human perception and computational vision by proposing a novel classification scheme of volumetric primitives based on considerations

about viewpoint invariance and perceptual organisation; these primitives are called *geon*s (geometrical ions). The fundamental assumption of RBC is that geons can be differentiated on the basis of properties in a single 2D image that are easy to detect and relatively viewpoint invariant.[4]

The basis for the use of these so-called *non-accidental properties* lies in the work of Gestalt psychologists like [Wertheimer 23] and in more recent theoretical studies on perceptual organisation such as [Binford 71], [Lowe 85], [Witkin & Tenenbaum 85] and others, and can be summarised in two key points:

1. Primal access to categories is based on edges; colour, texture and brightness have only secondary role;

2. Certain properties of edges in a 2D image are taken by our visual system as strong evidence of the existence of the same properties among 3D edges.

The edge-based primal access is supported by some psychological experiments (see [Biederman 87], also for bibliography) and it is also justified by the fact that in most cases surface characteristics are a computationally less efficient route to object recognition [Barrow & Tenenbaum 81]. As Biederman put it, *"we may know that a chair has a particular colour/texture but it is only the volumetric description that provides efficient access to the mental representation of a chair"* [Biederman 87].

The second argument about the importance of non-accidentalness is much stronger: these properties are *"carriers of statistical information"* [Lowe 85] and are very unlikely to arise due to accidental viewpoint or position. Lowe pointed out that there could be a theoretically infinite number of these relations (e.g., the aggregate of the properties on N edges) but only a small set of them are likely to be of any perceptual relevance. He found 10 distinct 2D image relations that can be considered non-accidental and from which 3D properties can be inferred.

With the task of differentiating volumetric primitives in his mind, Biederman built

[4] RBC has recently evolved into Geon Structural Description, or GSD [Hummel & Biederman 92]. As observed in [Tarr & Bulthoff 95], the fundamental difference is that in contrast to RBC, GDS-object based representations possess only *view-restricted* viewpoint invariance. GSD is viewpoint invariant only up to the visibility or occlusion of its parts, therefore resembling an *aspect graph* kind representation [Koenderink & vanDoorn 79].

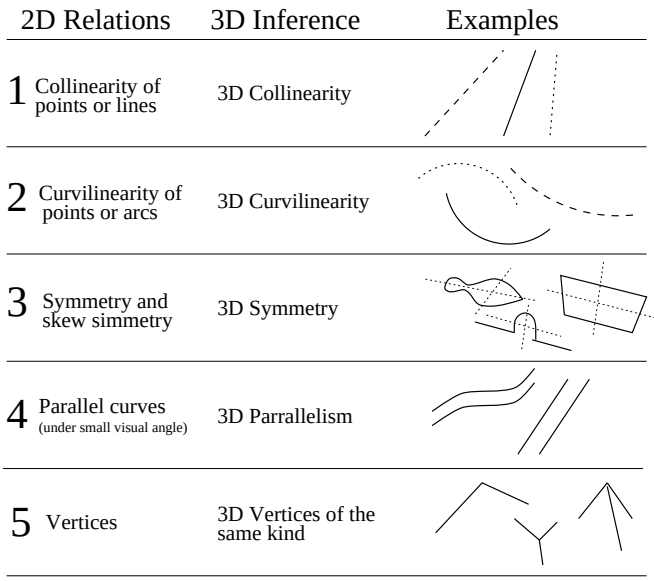


Figure 2.1: Set of five non-accidental properties used for 3D inference (redrawn from [Biederman 87])

upon Lowe's work and came up with 5 significant 2D relations and relative 3D inferences, shown in Fig. 2.1, that were then used as a perceptual basis for the generation of the set of basic primitives. He used generalised cylinders and considered four attributes: cross-section shape, symmetry and qualitative sweeping rule, and shape of axis. As he put it, *"from variations over only two or three levels in non-accidental properties of (those) four attributes of generalised cylinders, a set of 36 geons can be generated"*.

The four attributes that Biederman proposed are the following.

Cross-Section Shape. It is divided into "with straight edges" and "with curved edges". This seemingly simple classification is invariant over almost all viewpoints.

Cross-Section Symmetry. Three possibilities are considered: rotational & reflectional symmetry, reflectional symmetry and asymmetry. There is empirical psychological evidence that symmetrical shapes can be more quickly discriminated than asymmetrical stimuli and some degree of symmetry detection seems to be pre-attentively available [Garner 74, Checkosky & Whitlock 73]: our perception

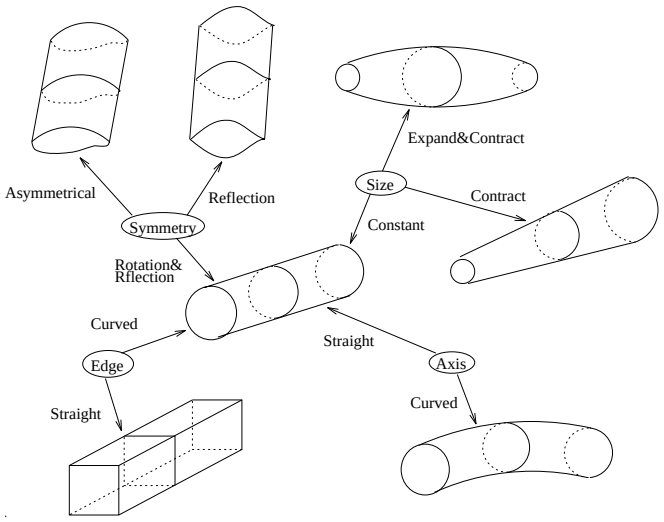


Figure 2.2: Example of taxonomy of a generalised cylinder based on the non-accidental attributes proposed by Biederman (redrawn from [Biederman 87]).

system seems to be biased towards symmetry [King *et al.* 76].

Cross-Section Sweeping Rule. Three variations are considered: constant, expanded and expanded & contracted. In the first case we have parallel sides, in the second non-parallel and in the third a lemon-shaped ellipsoid. As in the case of symmetry, there is empirical evidence (for example, Biederman cited [Ittleson 52]) that parallelism vs. non-parallelism is available pre-attentively and it is relatively (due to the pin-hole effect) viewpoint invariant.

Shape of Axis. As in the case of cross-section shape, it is divided into straight or curved and it is distinguishable by the non-accidental property of collinearity or curvilinearity.

The RBC theory can account for many object recognition capabilities but a more careful analysis leads to several issues that need to be considered for a full understanding of its scope and limitations; an excellent survey of these issues is given in [Dickinson *et al.* 93]. Some of the criticisms are discussed below.

Is the viewpoint independence argument valid? One of the main assumptions of the geon representation is that non-accidental properties of 2D images have

viewpoint invariance properties that allow one to infer 3D relationships. In a recent heated debate on this issue which appeared in Journal of Experimental Psychology [Tarr & Bulthoff 95, Biederman & Gerhardstein 95], Tarr *et al.* argued that the viewpoint invariant mechanisms advocated by Biederman lack *"generality for explaining a wide range of recognition phenomena"* and cannot even account for category recognition, exactly the problem that geon structural descriptions were introduced for; Tarr *et al.* in fact argue that there is enough evidence to support view-based mechanisms. Against this view, Biederman replied that *"geon structural descriptions provide a representation that distinguishes most entry- and subordinate-level classes and explains why complex objects are described as arrangements of viewpoint-invariant parts"* [Biederman & Gerhardstein 95].

Many objects cannot be properly represented by parts. This is certainly true. As pointed out by Pentland, this kind of representation is excellent for biological and man-made forms but *"becomes ill-suited when applied to complex or natural scenes like a mountain, a bush, a crumpled cloth, etc.: there is simply too much information in these kind of objects to be represented by simple models such as geons!"* [Pentland 86]. However, it is widely believed that the use of simple basic categories like geons helps perceptual systems achieve primal access to lawful relationships in the visual world.

Are there a limited number of primitives? This question is very tough to answer. In her classical work, [Rosch 78] supported the view that the use of a limited number of prototype parts to be combined and modified for describing things is common in human reasoning. Biederman noticed that, as much as for human speech where scientists have identified a limited number of building blocks (phonemes), RBC could well be one of the first attempts at finding building blocks for visual perception, although the complexity of vision tasks is far greater than that of speech understanding [Biederman 87]. Moreover, the intentional set of primitives proposed by Biederman has the interesting property of having minimal viewpoint uncertainty; although this feature cannot probably be used to classify all possible parts, certainly it provides an good start-point for

future attempts in defining a more general set of parts.

How to deal with structural details? This is a major problem that should be faced before trying to build any geon-based recognition system. As confessed by Dickinson in [Dickinson *et al.* 93], while experimenting with the OPTICA system, they had difficulties in *finding* suitable geon-based objects to test the system on because the texture, markings, flanges, ridges and so forth found on real objects make the task of recovering geons of paramount practical difficulty. Details do not define the coarse shape of objects and their parts, but they greatly affect segmentation and shape inference performance; Biederman and others that followed pure contour-based approaches did not address this key issue. Future research should account for structural detail if we aim at systems capable of working with real imagery.

Can geons be used to answer the "where" question? A vision system should not only identify objects but it should also be able to indicate their position in the space and their relative size. Pure geon-based techniques may be good for the identification task but are lacking in the localisation and characterisation tasks. There has been some research into ways of combining pure qualitative recognition with quantitative techniques to give geon-based systems full visual capabilities. An interesting solution is the one presented in [Metaxas *et al.* 93] in which the "what" and "where" are kept well separated by interfacing Dickinson *et al.*'s OPTICA system with a module that uses qualitative shape to constrain the fitting of physics-based deformable models (superquadrics) to the image. As a result of the fitting, they can have ready information about the position and the size of the objects and its parts.

Can we recover geons from real 2D imagery? At the moment, a honest answer is *NO*. The three works that address this problem are, thus far, [Dickinson *et al.* 92b], [Zerroug & Nevatia 94] and this thesis, all following different strategies but making a great deal of assumptions which are not always verified in real uncontrolled imagery. In principle, the recovery of geons should be easier than other visual tasks because it involves the exploitation of simple qualitative properties; however it is not yet clear what the correct approach is.

2.4 Relevant works in generic part recognition

There have been many works that claimed to perform part segmentation from 2D images, which can roughly be subdivided in *symmetry axes or skeleton techniques* (e.g., [Zerroug & Nevatia 94, Rom & Medioni 93]), *contour-based techniques* (e.g. [Freeman 78, Bennamoun & Boashash 94, Siddiqi & Kimia 95]), *primitive-based* (e.g., [Pentland 90, Hara *et al.* 92]), *graph-theoretic methods* (e.g., [Shapiro & Haralick 79]) and, less directly, *scale-space approaches* ([Lindeberg 94]). It is not within the scope of this section to give account for all this research but an attempt will be made in the more specific review given in Section 4.2.

In my opinion, however, three works stand out for elegance, significance and for their decisive departure from previous clichés. They will be extensively reviewed here; Table 2.1 summarises these three approaches and how they relate to this thesis.

Work	Approach	Main Assumption	Real Images?
This thesis	*Model-guided*	*Parts representable by generic shape models*	*Yes*
Dickinson	*Aspect-based*	*Region segmentation possible*	*Yes*
Bergevine	*Non-accidental features*	*Complete line-drawing*	*No*
Hummel	*Neural network*	*Detectable key features*	*No*

Table 2.1: Comparative table of the three reviewed approaches and how they relate to this thesis.

2.4.1 Bergevine and Levine's PARVO system

The PARVO system [Bergevin 90, Bergevin & Levine 93] was the first to explicitly address the recognition of geons from 2D images and was intended to literally apply the principles set forth by [Biederman 87].

It uses as input a manually-created line drawing of isolated multi-part objects. The first processing stage is to segment the line-drawing into individual geons by finding T-junctions pairs [Hoffman & Richards 85] occurring in the outline of the objects. At

this point all lines are labelled as belonging to certain parts. Lines are then grouped into faces that are later used to infer the class of each geon. Geons are then classified by the non-accidental features suggested by Biederman. The proposed scheme represents a complete object by an attributed graph in which each node represents a geon and connections define types of relations and relative sizes between geons. A pre-compiled object database is used with cross-indexing such that objects containing particular geons are listed together. The graph derived from the image is matched with one from the object database by mean of a weighted sum of the matching attributes.

The PARVO system was designed to give computational support and assess competence and performance of Biederman's ideas. So, there is no wonder in noticing how precisely all Biederman's indications have been followed. As a consequence, among the systems reviewed here, this is the one that most relies on idealised data; the system robustness is quite high for slight variations in line drawing properties (e.g., parallelism between two lines) and for image viewpoint transformations, but it falls badly apart when even a few key features (such as a single T-junction) are missing. For these reasons, its use with real imagery seems rather improbable.

2.4.2 Dickinson *et al.*'s OPTICA system

The approach proposed by Dickinson *et al.* [Dickinson *et al.* 92b] for the detection of volumetric primitives from 2D images is rather elegant and, apparently, efficient. The foundation of the work is the use of a 2D viewer-centred representation and aspect matching. They have implemented their ideas in a system called OPTICA (Object recognition using Probabilistic Three-dimensional Interpretation of Component Aspects) to demonstrate the validity of the approach, which is described in the following.

The fundamental stage is the construction of the so-called *aspect hierarchy*, which is exemplified in Fig. 2.3. First they choose a set of geon-like objects and then, by using a CAD system, they determined (by rotating the objects about the axes step-by-step) the set of all their viewer-centred 2D aspects, whose number is fixed and independent of the size of the object database. The aspects are represented by a 3-level hierarchy of 2D features (Fig. 2.3) that include, starting from the bottom level, groups of boundaries, faces and groups of faces, respectively. Then, the relations between these features are

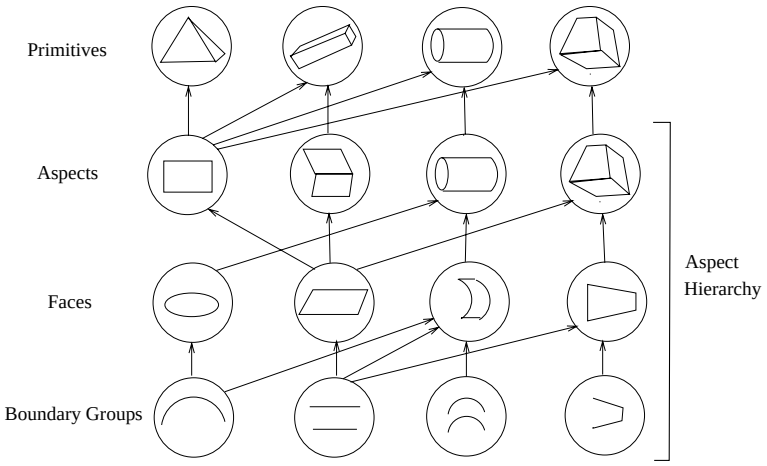


Figure 2.3: Dickinson*et al.*'s approach: The aspect hierarchy and the primitive level (redrawn from [Dickinson *et al.* 92a])

assessed from all view points in order to create a matrix of conditional probabilities of seeing each feature as a function of less complex 2D features. This matrix tells, for instance, what is the probability of viewing just two rectangles when seeing a block. This probability information is used to avoid combinatorial explosion in the search process. All the computations of the aspect hierarchy and matrices are performed *off-line*.

The primitive recovery is performed as follows.

First, the raw input image is processed by a Canny edge detector followed by simple morphology operations and then a connected edge algorithm is applied to extract a set of contours; possible curved contours are partitioned at significant curvature discontinuities using Ramer's algorithm [Ramer 72] and later some points are discarded if the angle between two circular arcs fitted to the left and right neighbours is near 180^o. From the set of partitioned contours, a procedure based on the Minsky's left shoulder rule for searching for minimal cycles in a graph (also used in [Pilu & Mainardi 90]) is utilised to yield the face graph. Each face is then analysed to extract properties (parallelism, symmetry and curvilinearity) and a new graph representing these relations is produced.

Then the aspect matching phase can be divided in four steps:

Recover Faces: From the region segmentation of the image, faces are classified according to those in the aspect hierarchy. If a face is occluded, groups of boundaries (one level down in the hierarchy) are used.

Recover Aspects: Next, regions are partitioned into groups, each corresponding to an aspect of a primitive, by region labelling and probabilities are used to limit the search space by starting from *face seeds* that are most likely to be good ones.

Recover Primitives: Next, primitive recovery consists of mapping the 2D aspects to 3D primitives, and recovering the connections between primitives.

Object Recovery: The final stage, not properly regarded as part detection, is to recognise the object contained in the scene. Groups of primitives are used as indices to the object database.

Despite the claim that the system could work with real images, only synthetic images and a controlled real one were presented as examples. Probably the problem was the reliable extraction of a correctly-partitioned face graph.

Later, Metaxas *et al.* [Metaxas *et al.* 93] built upon the work and, besides providing a quantitative front-end to the system by fitting superquadrics, they also changed the way the face graph was extracted by performing region segmentation by Saint-Marc and Medioni's edge-preserving adaptive smoothing filter [Saint-Marc & Medioni 88] to the image. Apparently the performance of the system improved, since they also presented an experiment in which an uncontrolled real image of shadowed and occluded polyhedra was used.

OPTICA is a very elegant system. The general paradigm of the use of characteristic views is applied to simpler parts of the object and the criteria defining the parts are based on the well known principles of perceptual organisation; moreover the multiple-level aspect hierarchy makes the system in principle capable of dealing with occlusion and limited contour fragmentation in a rather homogeneous and clean way. However, some problems need to be pointed out.

Firstly, the system is based, at intermediate level, on recovering minimal cycles of edges, which implies good connectivity. The level to which the system can deal with

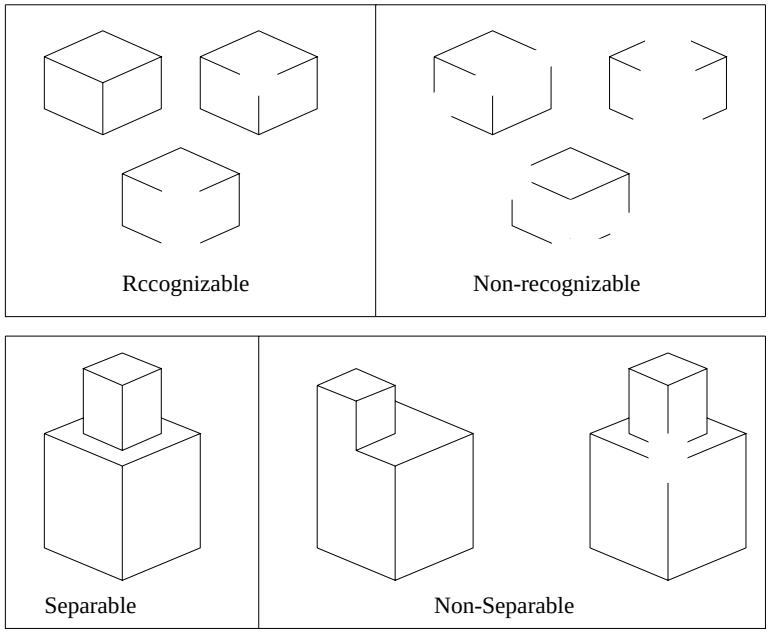


Figure 2.4: Problems with the Hummel and Biederman approach (redrawn from [Hummel & Biederman 92])

contour fragmentation is not well documented, since the proposed examples are either artificial or too simple. Arguably, a better face extraction algorithm could improve the reliability of the system but still no perfect face graph will ever be produced.

The second problem is more fundamental. The system performs well with objects whose constituting parts are separated by detectable and well defined edges. The mechanism for part segmentation relies in fact upon matching aspects directly into parts; if no edges separating parts are visible, a wrong mapping would occur. Many everyday objects have smoothed joints, with little intensity gradients (think of a doll, a telephone handset, etc) and this problem could hinder the use of the method.

2.4.3 Hummel and Biederman's approach

Hummel and Biederman presented the first (and thus far only) neural network approach to RBC [Hummel & Biederman 92]. Their prototype system aims at recognising isolated objects consisting of few geons out of eight geon classes from idealised line drawing data.

The network is composed of seven layers. The five lower ones are supposed to have fixed weights and to detect junctions, axes and blobs, geon features (axis curvature, etc) and geon unary relations (above/beside/below something, etc). The two upper layers are trainable to detect geon feature assemblies and objects. The capacity of extracting structural descriptions derives from the use of dynamic binding between cells. Specialised connections (FEL: Fast Enabling Links) in the first two layers parse images into geons by synchronising the oscillatory outputs of cells representing local image features (edges and vertices). Cells oscillate in phase if they represent features of the same geons and out of phase if they represent features of separate geons. In this scheme, object recognition corresponds to activating a particular cell of the seventh layer corresponding to a particular object. Because of the temporal conjoint of independent units, dynamic binding also allows a tremendous economy in the network structure.

The performance of the system in analysing line-drawings of assemblies of blocks is stunning. Some experiments for measuring the recognition time (such as for rotating the image) even seems to resemble that of humans, though Hummel and Biederman made no claims of biological plausibility of their implementation.

In their paper they provide some examples (reproduced here in Fig. 2.4) in which the recovery of geons or geon assemblies would be impossible. Apart from problems of this kind and over-grouping of too many features as reported by Hummel and Biederman themselves, it is very hard to believe that an approach like this could be adapted to real imagery in the foreseeable future.

Chapter 3

Modelling and Fitting Generic 2D Parts

In this chapter, I will discuss three different models that are used to qualitatively represent the outline of generic parts, and how they are fitted to point data sets. These three models are ellipses, deformable superellipses and statistical contour models.

The structure of the chapter is as follows. After an introduction, Section 3.2 presents an important achievement concerning the ellipse fitting problem, namely the first direct ellipse-specific least squares fitting method. Section 3.3 describes the deformable superellipse model (DSE) which is used to train, as presented in Section 3.4, a statistical part-like model that offers some advantages in terms of generality and ease of fitting over standard DSEs, as will be shown by some experiments. Finally, a summary and the contributions of the chapter are discussed.

3.1 Introduction

Ellipses, deformable superellipses and statistical contour models are proposed in this chapter as a coarse representation of the outline of generic parts.

These three models have each some advantages over the others depending upon the kind of data domain: I decided to include them all here because these three models were deemed simple enough to be easily fitted and yet with a sufficient representational power to coarsely represent outlines of a large class of parts of natural and man-made objects.

Of course, many other representations would be possible, such as Fourier contour models [Staib & Duncan 92], physically-based models [Pentland & Sclaroff 91], snakes [Kass *et al.* 88], higher order polynomials [Taubin 91], just to name a few. However, as models grow more and more complicated, they become more and more difficult to extract from images, a fact testified by the virtual absence of systems or techniques for fitting higher order part models to real and *unsegmented* images. This is hardly a surprise: high-order object and part models originate from the stream of computer graphics research, where modelling demands are somehow opposite to those of computer vision for many tasks such as grouping and recognition.

Simple low-order models have been suggested since the early days of vision research (e.g. [Binford 71] or [Marr & Nishihara 78]) as a good way of qualitatively modelling the essential structure of both solid and two-dimensional parts; there also is much psychological evidence [Biederman 87] that humans do use low-order models in early vision stages (see Sections 2.1 and 2.2 for a review).

The necessity of simplifying instead of complicating stems from the Gestalt school of thought [Wertheimer 23, Gibson 79] and, despite some criticisms (see, e.g., [Rock 83]), its validity has been extensively confirmed by the wealth of research in perceptual organisation. Simplicity can be exploited in a computational context when searching for significant relationships between image entities that are carriers of statistical information about the structure of the 3D space and the objects therein [Lowe 85] .

The choice of using simple but sufficiently powerful models reflects the need of extracting useful information from the natural scale of objects without being led astray by insignificant details: low-order models can be used to guide part-based perceptual grouping *towards the meaningful by discarding the meaningless.*

Last but not least, lay computational needs. Perceptual organisation is, as now commonly believed, a pre-attentive task that has huge computational requirements; if human-level perceptual organisation ability is ever to be achieved by computers, this is likely to be obtained only by operating on the essential structure first, according to the philosophy of [Pentland 86].

As we shall see in the rest of the thesis, the simple models discussed in this chapter suit

needs of simplicity, representativity and computational efficiency and this is precisely the reason for which they have been chosen for achieving part-based grouping in this work.

3.2 Ellipse fitting

The fitting of primitive models to image data is a basic task in pattern recognition and computer vision that reduces and simplifies the data to the benefit of higher level processing stages. Many models have been proposed and ellipses are undoubtedly amongst the most relevant, due to their simplicity and relative versatility, especially in the context of part-based recognition.

This section[1] deals with the fitting of ellipses to point data sets. The properties and the mathematics of ellipses are well known from elementary geometry and will not be reviewed; instead, a comprehensive review of former ellipse fitting methods is included as a prelude and justification of the new ellipse-specific fitting method presented in Section 3.2.2, which was introduced by [Fitzgibbon & Fisher 95] as a curiosity but the authors were unaware of its importance; I soon spotted its originality and significance, which was confirmed by extensive literature review, and set about providing the major contribution of a theoretical justification for it. Direct LSQ fitting of ellipses was, as we shall see, previously thought of as infeasible by leading researchers.

Several experiments to qualitatively assess performance and robustness of the method are presented along with some comments and proposal for future extensions. A quantitative analysis of the results can be found in [Fitzgibbon 96].

3.2.1 The LSQ ellipse fitting problem

Many techniques for fitting ellipses have appeared in the literature based on mapping sets of points to the parameter space, such as the Hough transform [Leavers 92] and accumulation methods [Rosin 93a]. These Hough-like techniques have some great advantages, notably high robustness and no need of pre-segmentation; however, they

[1] A shorter version of this section appears in [Pilu *et al.* 96a] and in [Fitzgibbon *et al.* 96] with different theoretical scaffolding and more quantitative experiments.

suffer from high computational complexity and non-uniqueness of solutions, which make them sometimes unusable for real applications.

We are concerned with the algebraic least squares fitting to scattered data of ellipses expressed by implicit polynomials. This is an important problem that, though well investigated, is still open to new solutions, both in term of accuracy, robustness and, last but not least, speed.

In the following, I give a succinct review of the most relevant works in ellipse fitting and its closely related problem, conic fitting. It will be shown that *the direct specific least squares fitting of ellipses had not yet been solved.*

Let us represent a generic conic as the zero set of an implicit second order polynomial:

$$F(\mathbf{a}, \mathbf{x}) = \mathbf{a}\mathbf{x} = ax^2 + bxy + cy^2 + dx + ey + f \tag{3.1}$$

where $\mathbf{a} = [\ a\ b\ c\ d\ e\ f\ ]$ and $\mathbf{x} = [x^2\ xy\ y^2\ x\ y\ 1\ ]^T$. $F(\mathbf{a}, \mathbf{x}_i)$ is called the "algebraic distance" of a point $\mathbf{x}_i$ to the conic $F(\mathbf{a}, \mathbf{x}) = 0$.

One way of fitting a conic is to minimise the algebraic distance over the set of N data points in the least squares sense, that is

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ \sum_{i=1}^{N} \left(F(\mathbf{a}, \mathbf{x}_i)\right)^2 \right\} \tag{3.2}$$

In order to avoid the trivial solution $\hat{\mathbf{a}} = \mathbf{0}_6$, the coefficients $\mathbf{a}$ of the polynomial are subject to a constraint. Linear constraints of the kind $\mathbf{c}\mathbf{a} = const$, where $\mathbf{c}$ is a constraint vector, lead to the solution of a system of N linear equations as in [Rosin 93b, Gander *et al.* 94]. Quadratic constraints of the form $\mathbf{a}^T\mathbf{C}\mathbf{a} = const$, where $\mathbf{C}$ is a *constraint matrix*, lead to a generalised eigenvalue problem.

Linear conic fitting methods have been investigated that use linear constraints that slightly bias conic fitting towards elliptical solutions. In particular, [Rosin & West 90] and [Gander *et al.* 94] investigated the constraint $a + c = 1$ and [Rosin 93b] $f = 1$, where he also analyses the pros and cons of the two normalisations $f = 1$ and $a + c = 1$ and shows that the former biases the fitting to have less eccentricity, therefore increasing the chances that the fit is elliptical.

Remarkably, Gander *et al.* recently published a paper entitled "Least Square Fitting

of Ellipse and Circles" [Gander *et al.* 94], in which the normalisation $a+c=1$ leads to an over-constrained system of N linear equations. The proposed normalisation is the same as the one in [Rosin & West 90, Porrill 90] (although they do not refer to them) and it does not force the fitting to be an ellipse, as we shall see in the experiments, or by just considering the hyperbola $3x^2 - 2y^2 = 0$, which satisfies the constraint $a+c=1$. It must be said, however, that in the paper they make no explicit claim that the algorithm is ellipse-specific.

In a seminal work, Bookstein's [Bookstein 79] showed that if a quadratic constraint is set on the parameters (e.g., to avoid the trivial solution $\mathbf{a} = \mathbf{0}_6$) the minimisation (3.2) can be solved by the rank-deficient generalised eigenvalue system:

$$\mathbf{D}^T\mathbf{D}\mathbf{a} = \mathbf{S}\mathbf{a} \ = \lambda\mathbf{C}\mathbf{a} \tag{3.3}$$

where $\mathbf{D} = [\mathbf{x}_1 \ \mathbf{x}_2 \cdots \ \mathbf{x}_n]^T$ is called *design matrix*, $\mathbf{S} = \mathbf{D}^T\mathbf{D}$ is called *scatter matrix* and $\mathbf{C}$ is called *constraint matrix*.

A simple constraint is $\|\mathbf{a}\| = 1$ but Bookstein used the algebraic invariant constraint $a^2 + \frac{1}{2}b^2 + c^2 = 1$, which leads to a rank-deficient generalised eigenvalue problem which he solves by block decomposition.

The Bookstein method could not constrain the fitting to be elliptical, in the sense that given some noisy elliptical data, the result could have well been an hyperbola or a parabola. However, it has been widely used in the past decade in most of the non-Hough based works that investigated, improved or used ellipse fitting.

In [Sampson 82], an improvement to the Bookstein method was presented that replaces the algebraic distance $F(\mathbf{a}, \mathbf{x})$ with a better approximation to the geometric distance

$$\frac{F(\mathbf{a}, \mathbf{x})}{\|\nabla_{\mathbf{x}} F(\mathbf{a}, \mathbf{x})\|} \tag{3.4}$$

which provided more stability and accuracy but unfortunately required an iterative algorithm. In [Taubin 91], the clever approximation of (3.2) using the distance (3.4)

$$\sum_{i=1}^{N} \left| \frac{F(\mathbf{a}, \mathbf{x}_i)}{\|\nabla_{\mathbf{x}} F(\mathbf{a}, \mathbf{x}_i)\|} \right|^2 \approx \frac{\sum_{i=1}^{N} |F(\mathbf{a}, \mathbf{x}_i)|^2}{\sum_{i=1}^{N} \|\nabla_{\mathbf{x}} F(\mathbf{a}, \mathbf{x}_i)\|^2}, \tag{3.5}$$

was proposed to turn the problem again into a generalised eigen-system, thereby allowing direct fitting.

In [Porrill 90] and [Ellis *et al.* 92], the Bookstein method was used to first fit a generic conic followed by a Kalman filter to iteratively minimise the distance (3.4), to gather new image evidence and to reject non-ellipse fits by testing the discriminant $b^2-4ac<0$ at each iteration. Porrill also gives nice examples of the confidence envelopes of the fittings. Rosin used either the Bookstein method alone [Rosin & West 90], (assuming that the shapes were elliptical) or again in conjunction with a Kalman filter as in [Rosin & West 95] and [Rosin 93b].

Some true ellipse-specific methods have been proposed that somehow "play" with the coefficients of the algebraic equation (3.1) to incorporate the constraint $b^2-4ac<0$ but all need iterative non-linear minimisation to be solved (see [Haralick & Shapiro 92], also for an extensive literature review).

Concluding, despite the amount of work, ellipse-specific direct fitting was left unsolved. If ellipses fitting was needed, one had to rely either on generic conic fitting or on iterative methods. Curiously enough, recently [Rosin & West 95] re-iterated this problem by stating that *ellipse-specific fitting is essentially a non-linear problem and iterative methods must be employed* for this purpose. In the following we show that this is no longer true.

In the rest of the section, I will refer for comparisons to Bookstein's [Bookstein 79], Gander's methods [Gander *et al.* 94] (although similar to [Rosin & West 90]) and to the algorithm that minimises Eqn. (3.5), which will be referred to as Taubin's method .

3.2.2 Direct least squares method

Let us consider a different quadratic constraint that corresponds to the well known quadratic algebraic invariant of a conic:

$$b^2-4ac=\mathbf{a}^T\begin{bmatrix}0&0&-2&0&0&0\\0&1&0&0&0&0\\-2&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}\mathbf{a}=\mathbf{a}^T\mathbf{C}\mathbf{a}<0 \tag{3.6}$$

This constraint was first introduced in [Fitzgibbon & Fisher 95] and it was shown to yield always elliptical solutions; the brief justification given there was that because of

the immateriality of the scale of $\mathbf{a}$, the inequality (3.6) can, without loss of generality, turned into $\mathbf{a}^T\mathbf{C}\mathbf{a} = -1$ and hence the minimisation (3.2) subject to the constraint (3.6) can again be formulated as in (3.3).

In the following, I give a theoretical account of the method by demonstrating its key feature of ellipse specificity, i.e. that *it gives always one and only one elliptical solution*; but before that, we need to state two Lemmas that will naturally lead to a uniqueness theorem[2].

Let $\mathbf{S} \in \Re_{n\times n}$ and $\mathbf{C} \in \Re_{n\times n}$ be symmetric matrices, with $\mathbf{S}$ positive definite. Let us define the *spectrum* $\sigma(\mathbf{S})$ as the set of eigenvalues of $\mathbf{S}$ and let $\sigma(\mathbf{S}, \mathbf{C})$ analogously be the set of generalised eigenvalues of (3.3).

Lemma 1 *The signs of the generalised eigenvalues of*

$$\mathbf{S}\mathbf{u} = \lambda\mathbf{C}\mathbf{u} \tag{3.7}$$

are the same as those of the matrix $\mathbf{C}$*, up to permutation of the indices.*

Proof: Let the *inertia* $i(\mathbf{S})$ be defined as the set of signs of $\sigma(\mathbf{C})$, and let $i(\mathbf{S}, \mathbf{C})$ analogously be the inertia of $\sigma(\mathbf{S}, \mathbf{C})$. Then, the lemma is equivalent to proving that $i(\mathbf{S}, \mathbf{C}) = i(\mathbf{C})$. As $\mathbf{S}$ is positive definite, it may be decomposed as $\mathbf{Q}^2$ for symmetric $\mathbf{Q}$, allowing us to write (3.7) as $\mathbf{Q}^2\mathbf{u} = \lambda\mathbf{C}\mathbf{u}$. Now, substituting $\mathbf{v} = \mathbf{Q}\mathbf{u}$ and pre-multiplying by $\mathbf{Q}^{-1}$ gives $\mathbf{v} = \lambda\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}^{-1}\mathbf{v}$ so that $\sigma(\mathbf{S}, \mathbf{C}) = \sigma(\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}^{-1})^{-1}$ and thus $i(\mathbf{S}, \mathbf{C}) = i(\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}^{-1})$. From Sylvester's Law of Inertia [Wilkinson 65] we have that for any symmetric $\mathbf{S}$ and nonsingular $\mathbf{X}$, $i(\mathbf{S}) = i(\mathbf{X}^T\mathbf{S}\mathbf{X})$. Therefore, substituting $\mathbf{X} = \mathbf{X}^T = \mathbf{Q}^{-1}$ we have $i(\mathbf{C}) = i(\mathbf{Q}^{-1}\mathbf{C}\mathbf{Q}^{-1}) = i(\mathbf{S}, \mathbf{C})$. □

Lemma 2 *If* $(\lambda_i,\ \mathbf{a}_i)$ *is a solution of the eigen-system (3.3), we have:*

$$sign(\lambda_i) = sign(\mathbf{a}_i^T\mathbf{C}\mathbf{a}_i).$$

Proof: By pre-multiplying by $\mathbf{a}_i^T$ both sides of (3.3) we have $\mathbf{a}_i^T\mathbf{S}\mathbf{a}_i = \lambda_i\mathbf{a}_i^T\mathbf{C}\mathbf{a}_i$. Since $\mathbf{S}$ is positive-definite, $\mathbf{a}_i^T\mathbf{S}\mathbf{a}_i > 0$ and therefore λ_i and the scalar $\mathbf{a}_i^T\mathbf{C}\mathbf{a}_i$ must

[2] I have not been able to locate any reference regarding the solution of generalised eigenvector problem with an indefinite constraint matrix.

have the same sign. □

Now we can state the following uniqueness theorem:

Theorem 1 *The solutions to the conic fitting problem given by the generalised eigen-system (3.3) subject to the constraint (3.6) include one and only one elliptical solution corresponding to the single negative generalised eigenvalue of (3.3). The solution is also invariant to rotation and translation of the data points.* [3]

Proof: Since the non-zero eigenvalues of $\mathbf{C}$ are $\sigma(\mathbf{C}) = \{ -2,\ 1,\ 2 \}$, from Lemma 1 we have that $\sigma(\mathbf{S}, \mathbf{C})$ has one and only one negative eigenvalue $\lambda_i < 0$, associated with a solution $\mathbf{a}_i$; then, by applying Lemma 2, the constraint $\mathbf{a}_i^T \mathbf{C} \mathbf{a}_i = b^2 - 4ac$ is negative and therefore $\mathbf{a}_i$ is a set of coefficients representing an ellipse. The constraint (3.6) is a conic invariant to Euclidean transformation and so is the solution (see [Bookstein 79]) □

Theorem 1 does not state anything about the *quality* of the unique elliptical solution, since classical optimisation theory states that it might not be the global minimum of (3.2) under our *non-positive definite* inequality constraint (i.e., the Kuhn-Tucker conditions are not verified [Wilkinson 65]). However, the physical solution (the actual ellipse) does not change under linear scaling of the coefficients and therefore it can be easily shown that the minimisation with the inequality constraint (3.6) can be equivalently turned to a minimisation with an equality constraint $\mathbf{a}^T \mathbf{C} \mathbf{a} = -1$. By doing so, as illustrated in [Fitzgibbon *et al.* 96], we can say that:

Corollary 1 *The unique elliptical solution is the one that minimises (3.2) subject to the constraint* $\mathbf{a}^T \mathbf{C} \mathbf{a} = -1$.

A more practical interpretation of this corollary is that the unique elliptical solution is a local minimiser of the *Rayleigh quotient* $\frac{\mathbf{a}^T \mathbf{S} \mathbf{a}}{\mathbf{a}^T \mathbf{C} \mathbf{a}}$ and thus the solution can also be seen as the *best least squares ellipse under a re-normalisation of the coefficients by* $b^2 - 4ac$. Although experimental evidence would suggest that this statement could be valid, a

[3] Since $\mathbf{C}$ is rank deficient, the eigen-system (3.3) should be solved by block decomposition like in[Bookstein 79]; however most numerical packages will handle this detail.

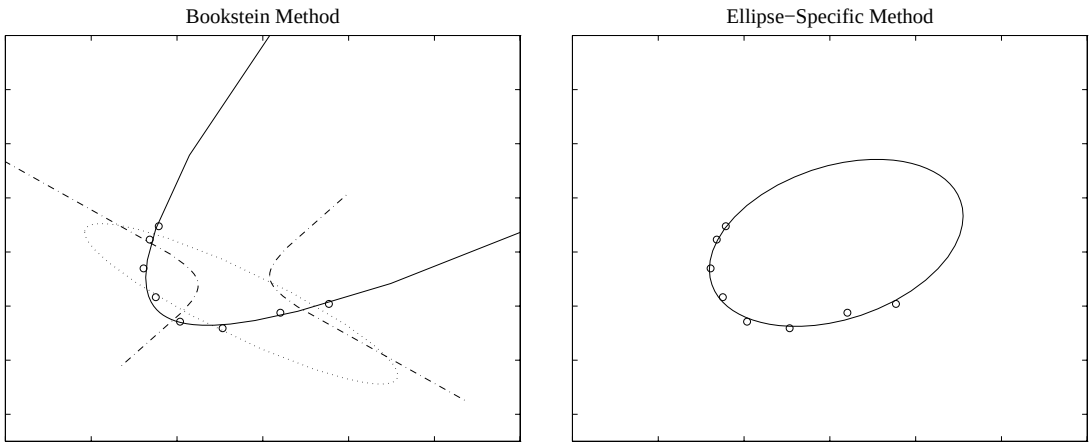


Figure 3.1: Specificity to ellipses. The left figure shows the three eigen-solution yielded by the Bookstein algorithm. The best LSQ fit is an hyperbola and the (incidentally) elliptical one is extremely poor. With the proposed ellipse-specific algorithm, the only solution satisfying the constraint is the best LSQ *elliptical* solution, shown on the right.

formal demonstration is currently not known to the author. This implicit normalisation turns singular for $b^2 - 4ac = 0$ and following the observations in [Rosin 93b], we can say that the minimisation tends to "pull" the solution away from singularities; in our case the singularity is a parabola and so the unique elliptical solution tends to be biased towards low eccentricity, which explains many of the following results, such as those in Figure 3.2. Curiously enough, in [Fitzgibbon *et al.* 96] we point out that since the discriminant $b^2 - 4ac$ is inversely proportional to the product of the radiuses, the minimisation (3.2) with implicit normalisation caused by our constraint acts much as the "minimum volume"[4] *heuristic* proposed in [Solina & Bajcsy 90], where the the non-linear minimisation was performed with an algebraic distance (similar to Eqn. (3.10)) that was multiplied by the product of the three radiuses. This provides a further intuitive feeling for the low-eccentricity bias of the proposed algorithm.

Finally, it ought to be said that, although experimental results show that our method performs better than others, it would be extremely hard to demonstrate theoretically that it is also better in terms of the *true geometrical distance* in the general case.

[4] Area, in the case of ellipses or superellipses.

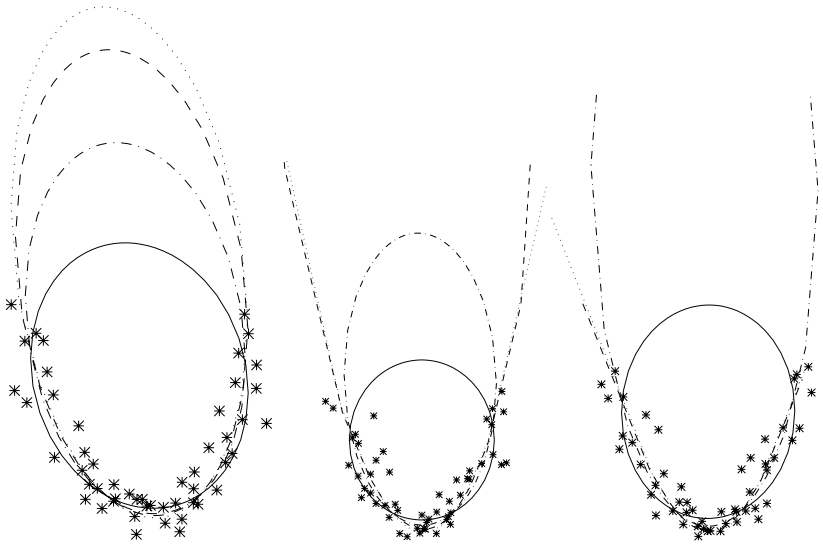


Figure 3.2: Fitting to noisy parabolic data. Encoding is Bookstein: dotted; Gander: dashed; Taubin: dash-dot; Ellipse-specific: solid. This example (after [Sampson 82]) shows the low-eccentricity bias of the ellipse-specific method, which is a desirable feature of an ellipse-fitting algorithm. See text for more details.

3.2.3 Experimental results and comments

This section gives some more experimental results that will help appreciate the goodness and the robustness of the proposed method.

First, let us now have a glimpse at what this ellipse-specificity means by using Figure 3.1 as an example. There, all the three generalised eigen-solutions of the Bookstein method are shown for the same set of data. The Bookstein algorithm gives an hyperbola as the best solution (solid line); the second best solution is, incidentally, an ellipse which is a very poor representation of the data. Conversely, the elliptical fit produced by the ellipse-specific method is a strikingly good one. Note that the solution produced by our method is not necessarily the best LSQ *conic* fit to the data, but the best LSQ *elliptical* fit.

Figure 3.2 shows three experiments designed after [Sampson 82] (who was inspired by [Gnanadesikan 77]) and which consist of the same parabolic data but with different realizations of added isotropic Gaussian noise ($\sigma = 10\%$ of data spread). In his paper, Sampson refined the poor initial fitting obtained with Bookstein algorithm by means of

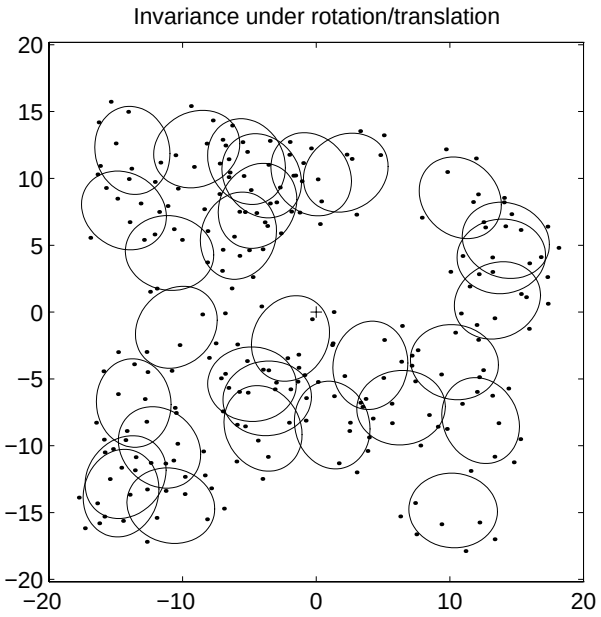


Figure 3.3: Invariance to Euclidean transformations. As expected from the invariance of the constraint, this experiment shows that the fitting results are unchanged up to the rotation and translation imposed on the data point set. See text for more details.

an iterative Kalman filter to minimise Eqn. (3.4). The final results were ellipses with low eccentricity that are qualitatively similar to those yielded by the proposed ellipse-specific direct method (solid lines) but at the *same computational cost of producing his initial estimate*.

As anticipated in the previous section, the low eccentricity bias of the ellipse-specific method is most evident in Figure 3.2 when compared to the results of the other methods, namely Bookstein (dotted), Taubin (dash-dots) and Gander (dashed). It must be again remarked that this is not surprising, because those methods are not ellipse-specific whereas the one presented here is.

The quadratic constraint that has been introduced not only bounds the fitted conics to be ellipses but it is also rotation and translation invariant. Figure 3.3 shows an experiment in which the fitting method was applied to several data point sets obtained by randomly rotating and translating an initial data set. The difference between expected and recovered semi-axes, centre positions and rotations were all zero up to machine precision.

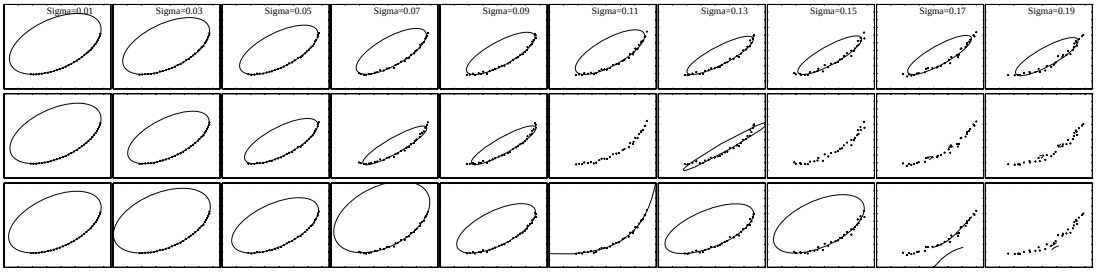


Figure 3.4: Stability experiments with increasing noise level (rightwards) of data spread. Top row: ellipse-specific method; Mid Row: Gander; Bottom Row: Taubin. The ellipse-specific method shows a much smoother and predictable decrease in in quality than the other methods. See text for more details. (In some figures the resulting conic is not drawn due to flaw in the conic drawing routine).

Let us now qualitatively illustrate the robustness of the ellipse-specific method as compared to the Gander and Taubin methods. A number of experiments have been carried out (of which here I present a couple, shown in Figure 3.4 and 3.5) by adding isotropic Gaussian noise to a synthetic elliptical arc; note that in both sets each column has the *same* set of points. More quantitative results can be found in [Fitzgibbon 96] and are not reported here for reasons of space.

Figure 3.4 shows the performance with respect to increasing noise level (see [Fitzgibbon & Fisher 95] for more experiments). The standard deviation of the noise varies from 3% in the leftmost column to 20% of data spread in the rightmost column; the noise has been set to a relatively high level because, as shown in the left-most column, the performance of the three algorithm is substantially the same at low noise level for *precise* elliptical data. The top row shows the results for the method proposed here. Although, as expected, the fitted ellipses shrink with increasing levels of high noise (as a limit the elliptical arc will look like a noisy line), it can be noticed that the ellipse dimension decreases smoothly with the increase of noise level: this is an indication of well-behaved fitting. This shrinking phenomenon is evident also with the other two methods but presents itself more erratically: in the case of the Taubin algorithm, the fitted ellipses are on average somewhat closer to the original one [Fitzgibbon & Fisher 95], but they are rather unpredictable and its ellipse non-specificity, as happens in the Gander case, sometimes yields unbounded hyperbolic fits.

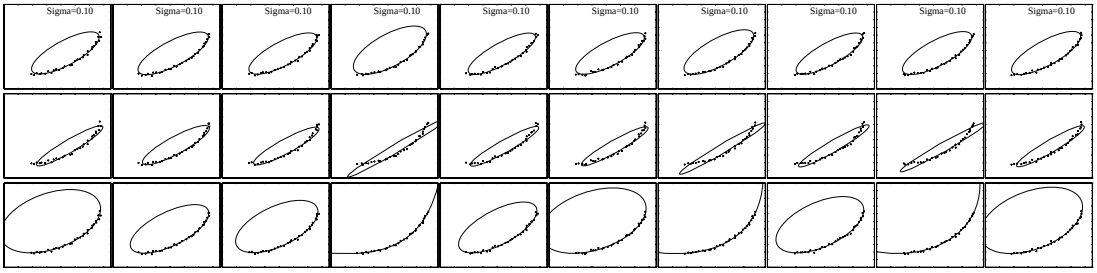


Figure 3.5: Stability experiments for different runs with same proposed method; Mid Row: Gander method; Bottom Row: Taubin method. The ellipse-specific method shows a remarkable robustness, as opposed to the other two. See text for more details.

The second set, shown in Figure 3.5, is concerned with assessing stability to different realizations of noise with the *same variance* ($\sigma = 0.1$). (It is very desirable that an algorithm's performance be affected only by the noise level, and not by a particular realization of the noise). This and similar experiments (see [Fitzgibbon *et al.* 96]) showed that our method has a remarkably greater stability to noise compared to Gander's and Taubin's.

Figure 3.6 shows some more examples with data set up by hand. The fittings of the method proposed here (solid line) are displayed along with Bookstein's (dotted) and Gander's (dashed). Data A is almost elliptical and indistinguishable fits were produced. Data B is elliptical too but more noisy; our fitting is clearly the best there. In C, Bookstein's fits to a hyperbola and in D and E so does Gander's. In F and G we have a "tilde" and two bent lines. Clearly these are not elliptic shapes but if data bounding were necessary, it can be seen that both Bookstein's and Gander's fail to do it, whereas our method nicely fits the data available and somehow delimits the region in which most data is enclosed.

Figure 3.7-left shows an *empirical* comparison between the number of FLOPS[5] versus data size N needed by Gander method, which involves the solution a system of N linear equations, and ours, which requires the solution of generalised eigenvalue problem. For a small N, the setting up of the scatter matrix dominates the execution time, whereas as N grows, (after about 150 data points in the example) the eigen-system method

[5] The number of FLOPS was determined using the MATLAB environment and it can be seen as an implementation independent estimate of the computational complexity.

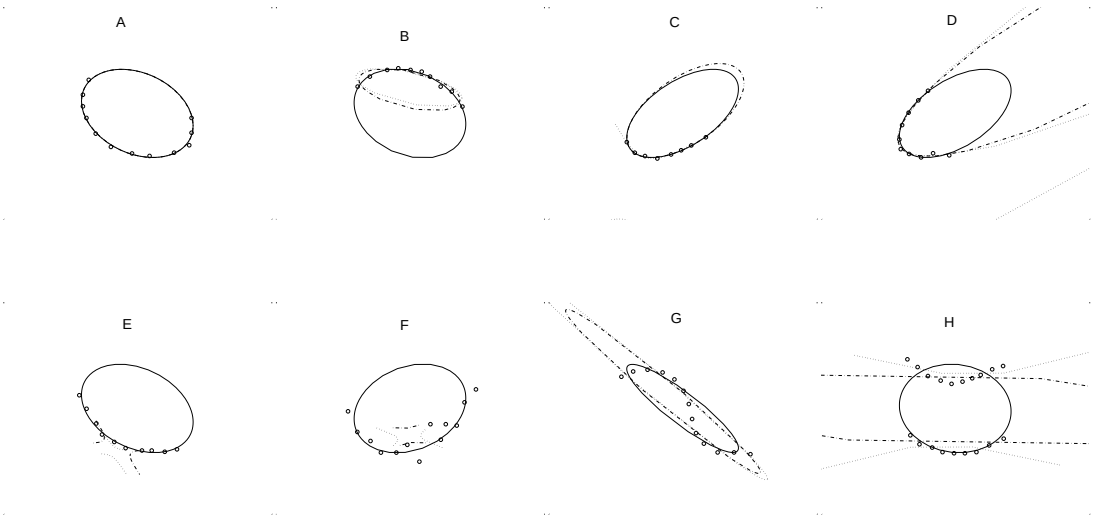


Figure 3.6: Fitting examples to hand-input data. Encoding is Bookstein: dotted; Gander: dashed; Taubin: dash-dot; Ellipse-specific: solid. The ellipse-specificity of the proposed method allow to bound the data points even in case where the pattern is far from elliptical. See text for more details.

becomes faster. More discussion about computational complexity can be found in [Fitzgibbon & Fisher 95].

Finally, Figure 3.7-right gives a simple six-line MATLAB implementation that, once again, shows the simplicity of the method. I have also implemented an on-line interactive JAVA demo of the method in which results can be compared with Bookstein and Taubin methods. The demo can be tried out with any Java-enabled Web browser at:

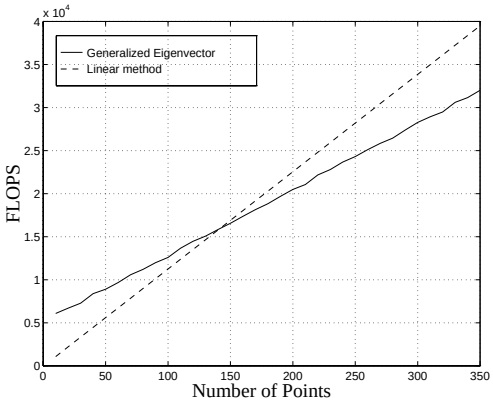
`http://vision.dai.ed.ac.uk/maurizp/ElliFitDemo/demo.html`

3.2.4 Summary and future work

This section has presented a least squares ellipse fitting method which is specific to ellipses and direct at the same time; other previous methods were either not ellipse-specific or iterative.

The method is possibly the best trade-off between speed and accuracy for ellipse fitting and its uniqueness property makes it also extremely robust to noise and usable in many applications, especially in industrial vision.

In [Fitzgibbon 96], a great deal of experiments and a better characterisation of the




```
% x,y are vectors of coordinates
function  a=fit_ellipse(x,y)
% Build design matrix
  D = [ x.*x x.*y y.*y x y ones(size(x)) ];
% Build scatter matrix
  S = D'*D;
% Build 6x6 constraint matrix
  C(6,6)=0; C(1,3)=-2; C(2,2)=1; C(3,1)=-2;
% Solve generalised eigen-system
  [gevec, geval] = eig(S,C);
% Find the only negative eigenvalue
  [NegR, NegC] = find(geval<0 & ~isinf(geval));
% Get fitted  parameters
  a = gevec(:,NegC);
```

Figure 3.7: Left: Empirical FLOP count comparison between linear and generalised eigenvector fitting methods. Right: Simple 6-line Matlab implementation of the proposed fitting method.

noise performance of the method are provided.

Three directions of further work can be identified. Firstly, further analysis of the new ellipse-specific direct least squares algorithm could be done in order to *theoretically* characterise its noise performance by using the eigenvalue perturbation theorem [Wilkinson 65]. Secondly, it would be nice to assess its benefits when used as a generator of initial estimates for iterative fitting methods such as [Sampson 82] and [Porrill 90]. Finally, a bias correction method should be studied as done in [Kanatani 94] to further increase performance (as also suggested by Kanatani himself in a personal communication on a draft of [Fitzgibbon *et al.* 96]).

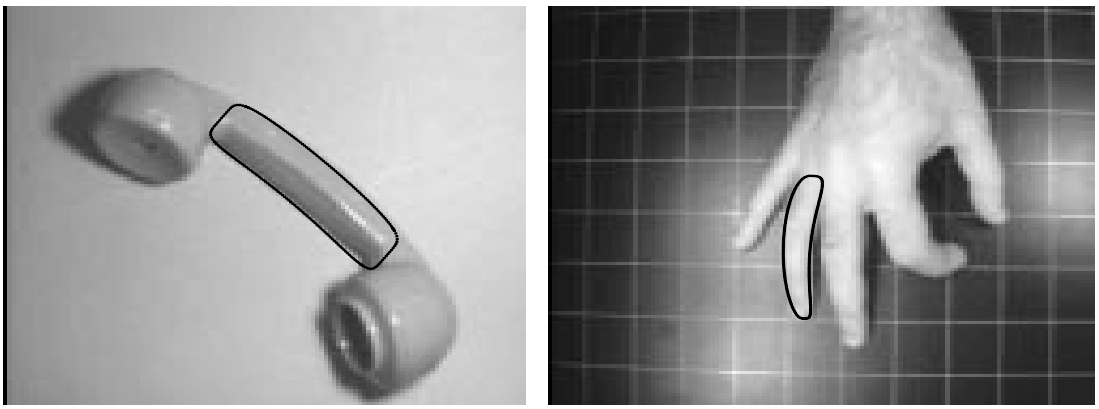


Figure 3.8: Two examples of modelling outline of parts by deformable superellipses.

3.3 The deformable superellipses model

Superellipses and their 3D extension superquadrics were introduced by the Danish designer Piet Hein [Gardner 65][6]. These curve and surface representations have been brought into the computer graphics and vision community by [Barr 81] and, in particular [Pentland 86], who used this new primitive to coarsely represent parts of objects very compactly. They can represent many closed 2D and 3D shapes (e.g. [Pentland 86, Pentland & Sclaroff 91] or [Raja & Jain 92a]) in a straightforward and natural way by using few parameters and moreover simple deformations can be applied to extend their modelling capabilities. Figure 3.8 shows two simple examples in which deformable superellipses are used to describe the outline of two simple parts.

The deformable superellipse model (DSE) was used in early stages of this work to represent part outlines but it has been dropped in favour of a properly trained Point Distribution Models as it will be shown in in Section 3.4. Some of the techniques that were devised for fitting DSEs are of a certain interest but they have not been included here because I deemed them not related to the rest of the work.

[6] Although he is normally referred as the inventor of superellipses, I have tracked their origin back to a study presented in 1818 by the French mathematician Gabriel Lamé on the curves of the form $(x/a)^n + (y/a)^n = 1$, which include superellipses.

A superellipse can be described in parametric form by:

$$\begin{cases} x = f_x(\theta) = a_x cos(\theta)^\epsilon \\ y = f_y(\theta) = a_y sin(\theta)^\epsilon \end{cases} \qquad -\pi \leq \theta \leq \pi \tag{3.8}$$

where a_x and a_y are the two semi-axes and $0 \leq \epsilon \leq 1$ is the roundness parameter. By eliminating θ, its implicit equation is:

$$\left(\frac{x}{a_x}\right)^{2/\epsilon} + \left(\frac{y}{a_y}\right)^{2/\epsilon} = 1 \tag{3.9}$$

The function:

$$F(x, y, a_x, a_y, \epsilon) = \left(\frac{x}{a_x}\right)^{2/\epsilon} + \left(\frac{y}{a_y}\right)^{2/\epsilon} \tag{3.10}$$

is called the "inside-outside" function, as its value determines whether a given point (x, y) lies inside, right on or outside the superellipse contour:

$$F(x, y, a_x, a_y, \epsilon) = \begin{cases} > 1 \ : \ Outside \\ = 1 \ : \ On\ the\ contour \\ < 1 \ : \ Inside \end{cases}$$

Either simple or complicated deformations can be applied to the basic superellipse shapes.

For the sake of self-containedness, below I give the mathematical description of the two simple deformations used in this work, linear tapering and bending, which have been derived from the 3D case in [Solina & Bajcsy 90].

Let the superellipse shape $\mathbf{S}$ be expressed in term of its vectors of coordinates $\mathbf{x}$ and $\mathbf{y}$ and let $\mathbf{X}$ and $\mathbf{Y}$ be the coordinates after the deformations.

Linear Tapering: The tapering deformation along the y axis is defined as:

$$Taper(K,\ \mathbf{S}) = \begin{cases} \mathbf{X} = g_x(y)\mathbf{x} \\ \mathbf{Y} = \mathbf{y} \end{cases} \tag{3.11}$$

If $g_x(y)$ is linear the tapering will also be linear. By setting $g_x(y) = \frac{K}{a2} + 1$, with $-1 \leq K \leq 1$, we have linear tapering ranging from increasing ($K > 0$), constant ($K = 0$) to decreasing cross-section ($K < 0$).

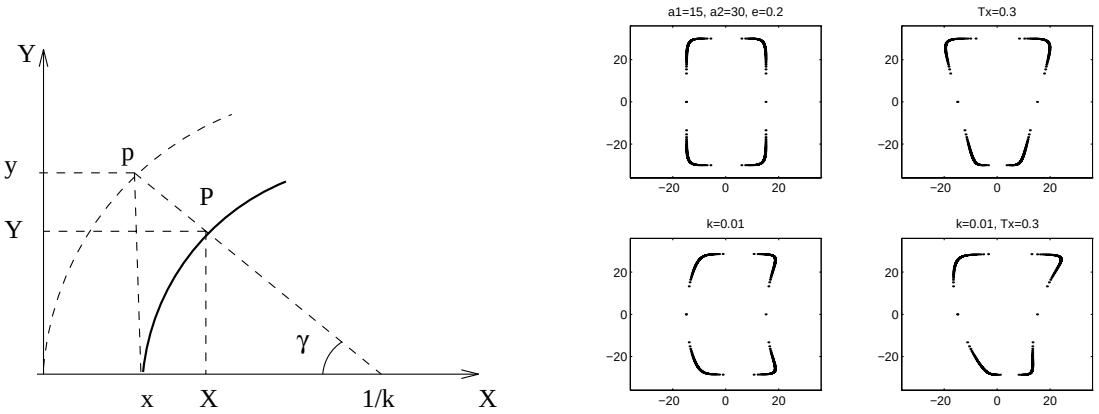


Figure 3.9: Bending Geometry Setting (left) and some examples of DSE (right) sampled linearly (see Appendix A).

Circular Bending: Figure 3.9 (left) sketches the geometry of the circular bending; only one parameter is needed to describe this deformation. As shown in the figure, the bending is applied along the y axis in the positive x direction. p is the original point position and P is the position when the deformation is applied. The deformation is given by:

$$Bend(b,\ \mathbf{S}) \begin{cases} \mathbf{X} = \mathbf{x} + sign(b) * (\sqrt{\mathbf{y}^2 + r^2} - r) \\ \mathbf{Y} = sin(\gamma) * r \end{cases} \tag{3.12}$$

where:

$$\begin{aligned} R &= a_y/|b| \\ r &= R - |\mathbf{x}| \\ \gamma &= atan(\mathbf{y}/r) \end{aligned}$$

and $-1 \leq b \leq 1$ is the bending control parameter. Differently from [Solina & Bajcsy 90], here the bending parameter is normalised to a_y and bending in both directions has been introduced.

Figure 3.9 (right) shows four superellipses, without deformation (top-left) with linear tapering (top-right), with bending (bottom-left) and with a combination of them (bottom-right). Note that in the examples, the parameter θ in Eqn. 3.8 is sampled linearly, causing a remarkably uneven sampling distribution. Solutions to this problem have been proposed in Franklin [Franklin & Barr 81] and also in [Pilu & Fisher 95].

A combination of deformations should be carried out by doing first the deformations that are more shape preserving (see e.g. [Leyton 92] or [Solina & Bajcsy 90]). In our case, the right order is taper first and bend afterwards. The complete transformation chain that brings a natural superellipse $\mathbf{S}$ in canonical position into the deformed shape $\mathbf{S}'$ in the image reference is therefore:

$$\mathbf{S}' = Trans(t_x, t_y, Rot(\theta_{opt}, Bend(b, Taper(K, \mathbf{S}))))$$

where p_x, p_y and θ_{opt} are the translations and the rotation, respectively.

3.4 Representing generic parts by a PDM

In the previous two sections, ellipses and superellipses were suggested as coarse part models and a new ellipse fitting method was presented. Superellipses have, however proven rather awkward to fit to scattered data.

This section[7] addresses the following problem: How can we make a complicated mathematical shape model simpler and easier to fit while keeping a comparable representational power? The proposed solution is to use the original model itself – which represents a class of shapes – to train a Point Distribution Model (PDM). This model will be extensively used in the two chapters that follow to coarsely represent the outlines of parts of objects.

Firstly, I give a description of PDMs and then show how the training set is built from deformable superellipse models and give some examples of parametric shapes thus obtained. Then, the fitting procedures to point data sets is presented along with some experiments. The section concludes with some comments and hints for future work.

3.4.1 The Point Distribution Model

Point Distribution Model (PDM) is a term coined by Cootes *et al.* [Cootes *et al.* 91] to indicate statistical finite-element models built from a *training set* of labelled contour landmarks of a large number of shape examples. The method has received lots of attention recently because of its flexibility and generality.

Let us indicate by Σ_2^n [Kendall 84] the set of shapes defined by a labelled set of n two-dimensional points $P_i = (x_i, y_i)$, also called *landmarks*. We desire to model a certain class of similar shapes belonging to Σ_2^n in order to identify and parametrise their significant degrees of freedom.

A well known tool for achieving this dimensionality reduction is the Karhunen-Loeve transform, or Principal Components Analysis (PCA) [Joliffe 86], by which a relatively large set of examples is used to infer global statistical properties of the whole set.

From a set of examples, n landmark points are chosen, labelled and put in corre-

[7] A shorter version of this work appears in [Pilu *et al.* 96b].

spondence across the whole training set. Let[8] $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_s}$ be the N_s aligned shape examples, each represented by $2n$-long vectors of landmark coordinates:

$$\mathbf{x}_i = [x_{i,1} \quad y_{i,1} \quad x_{i,2} \quad y_{i,2} \quad \cdots \quad x_{i,n} \quad y_{i,n}]^T.$$

The mean shape is calculated by averaging the position each coordinate point, that is

$$\overline{\mathbf{x}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbf{x}_i$$

and the $2n \times 2n$ (positive definite) covariance matrix of the points is given by

$$\Lambda = \frac{1}{N_s} \sum_{i=1}^{N_s} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$$

Let $(\lambda_1, \mathbf{p}_1), (\lambda_2, \mathbf{p}_2), \ldots, (\lambda_{2n}, \mathbf{p}_{2n})$ be the eigenvalue-eigenvector pairs of Λ sorted such that $\lambda_i \geq \lambda_{i+1}$. As is well known from statistics, the physical meaning of the eigenvector of a covariance matrix is a hyper-direction ($2n-$dimensional in our case) along which normal the variance of the point distribution equals the corresponding eigenvalue. Therefore the eigenvectors corresponding to the largest eigenvalues most describe the statistics of the point distribution.

This property of the eigenvalue decomposition is the key that has been cleverly used in [Cootes *et al.* 91] for *approximating* any shape $\mathbf{x}$ in the training set by a weighted sum of displacements in the direction of the t most significant eigenvectors with respect to the mean shape, that is:

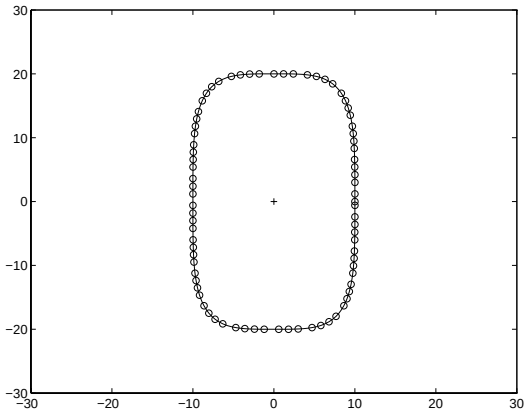
$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b}, \tag{3.13}$$

where $\mathbf{P} = [\mathbf{p}_1 | \mathbf{p}_2 | \cdots | \mathbf{p}_t]$ and the weights $\mathbf{b} = [b_1 b_2 \ldots b_t]$ are called *modes of variations.*

Equation (3.13) not only allows to represent the training set but represents *de facto* a parametric model of the class of training shapes and hence allows to generate new shapes in the class, provided that the b_is are kept within proper ranges.

By the nature of the decomposition used, each λ_i is the variance of the corresponding b_i over the training set and therefore the ranges for the b_i should fall within ± 2 or $3\sqrt{\lambda_i}$ [Cootes *et al.* 91].

[8] In the following we shall use the notation of [Cootes *et al.* 91].



	λ_i	η_i
Mode 1	5692.0	56.9 %
Mode 2	1933.3	19.8 %
Mode 3	1509.9	15.3 %
Mode 4	615.2	6.2 %
Mode 5	91.5	0.9 %
Mode 6	50.1	0.5 %
Mode 7	48.3	0.4 %

Figure 3.10: Landmarks of the natural superellipse model (left) and contribution of each mode to the overall point variance over the training set (right).

3.4.2 Building the training set

A properly built PDM can well represent the kind of variability wanted for modelling shapes like DSE in terms of dimension, bending and tapering, squareness and also shearing; however a method for efficiently building a large set of samples had to be devised and the most natural one was to *use the DSE model to train the PDM.*

A number ($N_s = 2000$) of random superellipses were generated, their contours subsampled at equal distance by the method proposed in [Pilu & Fisher 95] and, from the set of points, the PDM was built as in the previous section. Figure 3.10-left shows a natural superellipse in canonical position with the landmark points; we used for the experiments $n = 80$, i.e. 20 landmarks in each DSE quadrant, which has been found a good compromise that also avoided point interpolation in the fitting phase as done, e.g., in [Hill & Taylor 92]. By using the DSE construction as given in Sec. 3.3, the range of the parameters used to generate the random training set was $-5 \leq a_x \leq 5$, $-10 \leq a_y \leq 10$, $-0.7 \leq K \leq 0.7$, $-0.7 \leq b \leq 0.7$ and, finally, $0.1 \leq \epsilon \leq 0.9$. Note that the absolute values of a_x and a_y is irrelevant, since the PDM will be rescaled during the fitting stage [Cootes & Taylor 92]; their ratio, however, expresses the the eccentricity of the shape, which is usually assumed elongated along the y axis.

Table 3.10 shows the contribution in percentage of the first seven modes to the total

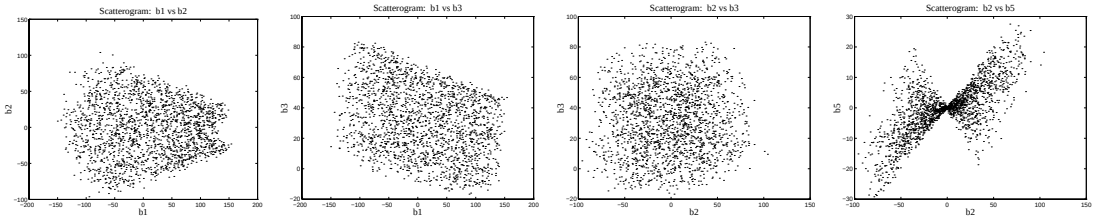


Figure 3.11: Four examples of scattergrams of the modes of variation. Low-order modes are relatively uncorrelated with each other whereas more and more correlation is found for higher order modes.

variance of the training set [Joliffe 86] given by:

$$\eta_i = \frac{\lambda_i}{\sum_{j=1}^{2n} \lambda_j}.$$

It can be seen that most of the shape variations are covered by the first 4 modes which, as we shall see in a moment, are strictly related to the actual deformable superellipse parameters.

The principal component analysis method teaches us to use scattergrams to check correlation between the modes over the training set: a scattergrams of two modes should look like a cloud of random points if they are uncorrelated with each other [Joliffe 86].

Figure 3.11 shows four scattergrams of various modes parameters computed over the training set. In our case the first 3 modes (b_1 vs b_2, b_1 vs b_3 and b_2 vs b_3 in Figure 3.11) look relatively uncorrelated but not for higher order modes such as, e.g. b_2 vs b_5.

An interesting experiment is to relate the original deformable superellipse parameters – used for building up the training set – to the modes of variation in order to assess their reciprocal correlations. Figure 3.12 shows the scattergrams of the first seven modes $b_1, ...b_7$ (rows) with respect to the five deformable superellipse parameters (see Sec. 3.3) a_1, a_2, ϵ ("e" in the figures), K and b are represented in the columns; a marked line-like pattern in the scattergram indicates strong correlations.

It can be seen that modes b_1, b_2, b_3 and b_4, chiefly correlate with a_2, b, a_1 and K, respectively, whereas they are pretty much uncorrelated with other parameters. This is a very welcomed behaviour, because it allows easy and natural classification of the

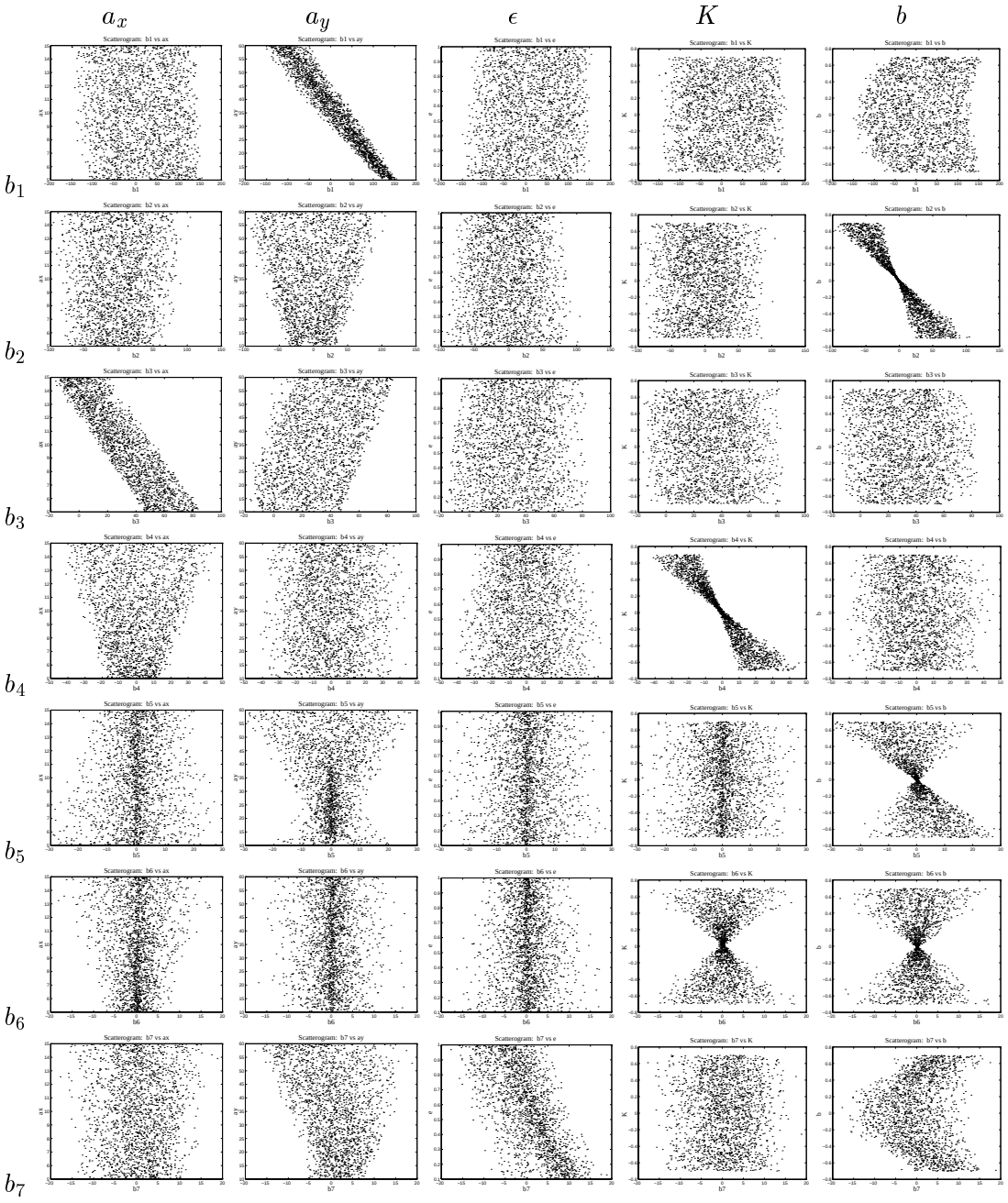


Figure 3.12: Scattergrams that the relates the modes of variation to the original deformable superellipse parameters over the training set. High concentration of points around a line indicate high correlation. It can be seen that modes b_1, b_2, b_3 and b_4, chiefly correlate with a_2, b, a_1 and K, respectively and pretty much uncorrelated with other parameters.

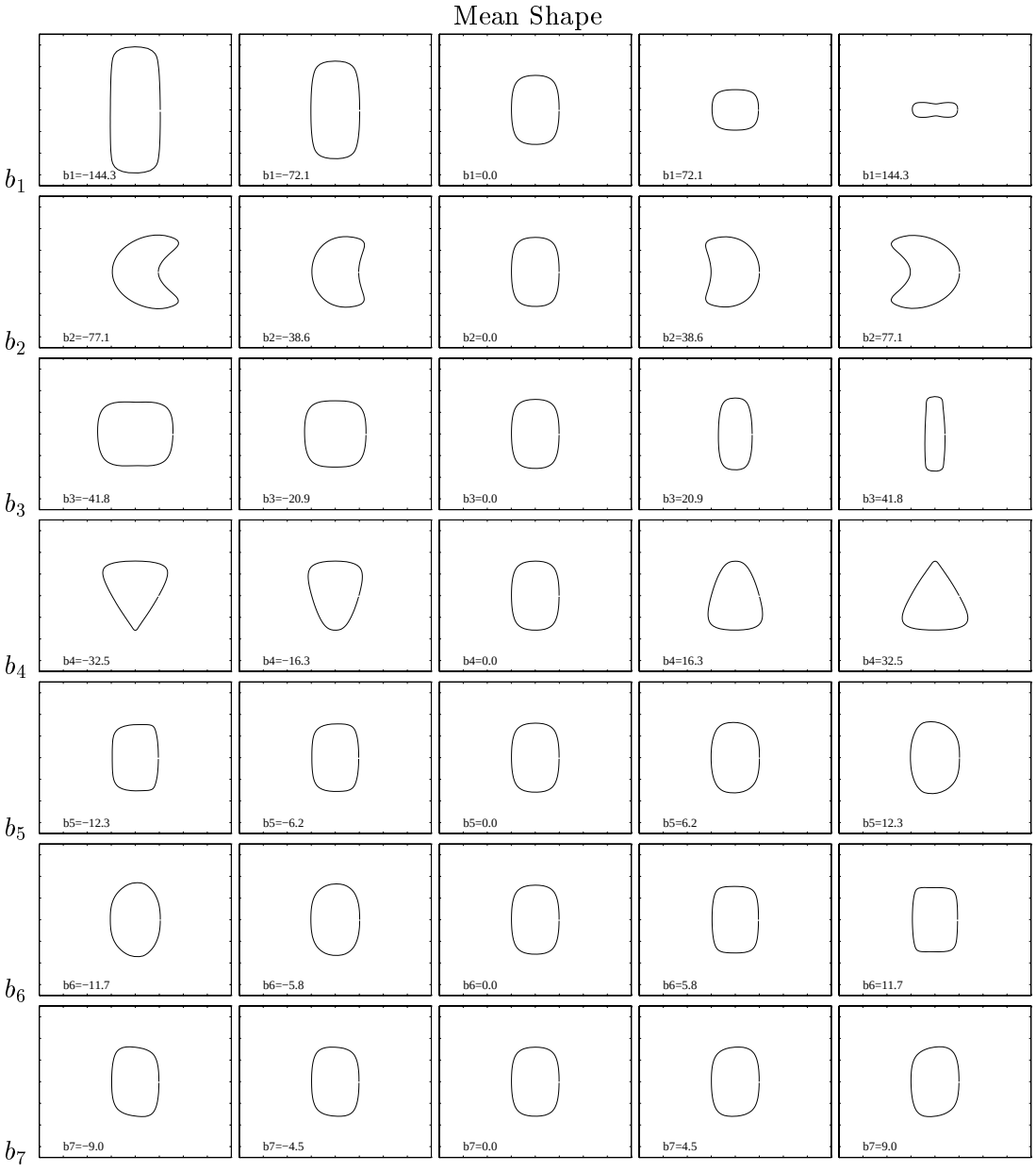


Figure 3.13: Parametrisation of the PDM model. The modes of variation control the actual PDM shape in a rather neat way. The first four modes directly control vertical height, bend, width and tapering, respectively, whereas the last three produce, in combination and unexpectedly, slight horizontal tapering, squaring and shearing.

shapes by just using modes straight away. The strong non-linearity of the roundness deformation (controlled by ϵ), which actually does not involve major structural changes in the shape, is not neatly correlated with any mode, although slightly with b_6. The roundness deformation is strongly non-linear and therefore this behaviour was somewhat expected.

This correlation between modes and shapes comes to the fore in Figure 3.13, in which the PDMs are shown for five different values of the first seven modes, one per row. The first four modes neatly control single shape features of vertical height, bend, width and tapering, whereas the last three produce, in combination and unexpectedly, slight horizontal tapering, squaring and shearing, which also nicely enhances the model's representational power.

Interestingly, there is a suggestive comparison of these results with Leyton's causal theory of shape [Leyton 92], which proposes a natural order with which shapes are deformed. The contributions η_i of each mode to the overall shape variance indicate that the most influential shape factors are, in order of importance, the major axis length, bending, shape width and, finally, tapering; this is remarkably in accordance with Leyton's theory, although at the moment it can be seen as a mere speculation.

3.4.3 Prelude to fitting: Initialisation

Before leaping to the description of the fitting of PDMs to point set data, this subsection briefly describes the initialisation stage that has been employed in this thesis for fitting PDMs to point data sets.

The non-linear optimisation needed for fitting shapes to data is a non-trivial one and the most important factor is to find a good initial position of the model. The methods employed to determine this initial state vary according to the kind of data available, how complete and noisy this data is expected to be, and so forth. In most (if not all) works on parametric model fitting, the initialisation is performed semi-automatically (e.g. [Lowe 91]) or the fitting is performed on an single isolated part such as in [Solina & Bajcsy 90]. (A famous exception is [Metaxas *et al.* 93], as reviewed in Sec. 2.4).

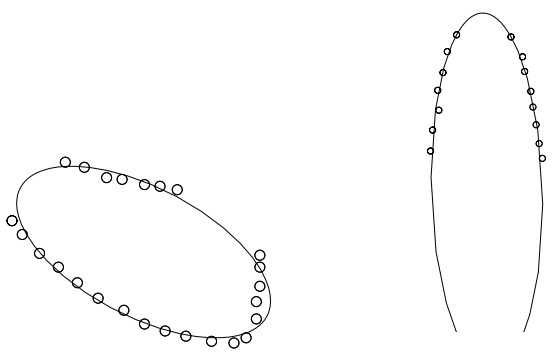


Figure 3.14: Initialisation by ellipse fitting: successful (left) and unsuccessful (right). In the first case the fitted ellipse nicely represents "part-like" the data points. In the second case, the two sides of a likely part have been fitted by a exceedingly large ellipse.

In this thesis, the initialisation from real images is performed by forming small groups of codons (currently pairs; see next chapter). However, since codons can be regarded as sets of points, we shall briefly discuss here two methods for producing coarse initialisation from point data sets, along with their advantages and drawbacks.

The two techniques in question are the computation of moments and the initialisation by ellipse fitting. Both these methods have been used in the literature, such as in [Solina & Bajcsy 90] and [Borges 96, Wu & Levine 94] in the case of the fitting of superquadrics to range data. However in most works, the surface data was supposed *structurally* complete (i.e. both main sides present) and initialisation problems with incomplete data sets were not even mentioned; on the other hand, [Leonardis *et al.* 94] employed a different strategy that used small surface patches to generate and grow superquadric hypotheses.

In the *moments of inertia* method (see, e.g., [Solina & Bajcsy 90]), the model position is found by computing the centre of mass of the points and the orientation by picking the eigenvector of the points' covariance matrix with the largest eigenvalue. Moments of inertia are intrinsically robust but do not always give a good indication of the object position, especially when the data available is incomplete (think of, e.g., half a circle).

Ellipses can also be used – especially in the context of part recognition – to estimate coarse model orientation and position from the set of points, with excellent results for

roundish shapes. However there are cases in which the ellipse fitting gives completely wrong results, the typical case being two convergent lines or curves, such as shown in Figure 3.14.

Throughout this work I have performed initialisation by ellipse fitting because I found it much more suitable for coping with substantial missing contour portion. In this context, the ellipse specificity of the new ellipse fitting method presented in Sec. 3.2 has been extremely useful (see the examples of Fig. 3.6). However, further work is needed in this respect.

The method used for transforming the implicit equation of a fitted ellipse – such as in Sec. 3.2 – into the explicit form that elicits axes length, orientation and the centre position, is to first find the rotation that annuls the xy term of the implicit conic equation and then the translation that annuls the terms in x and y. The method is better detailed in [Fitzgibbon 96].

3.4.4 Standard PDM fitting

In order to align a parametric PDM with image data, it is necessary to employ an iterative fitting strategy.

The method used here is essentially the one of Cootes *et al.*, which has been extensively experimented with on several occasions [Cootes & Taylor 92, Cootes *et al.* 94, Hill & Taylor 92]. For the sake of self-containedness, the procedure is succinctly described in the following.

At a certain iteration, let $\mathbf{X}$ be the vector of landmark points (built as in Sec. 3.4.2) in the *image reference frame* and let $\mathbf{x}$ be the same landmarks but in the *PDM reference frame*, i.e. that was used for building the training set. Let[9]

$$d\mathbf{X} = [\, dX_1 \;\; dY_1 \;\; dX_2 \;\; dY_2 \;\; \cdots \;\; dX_n \;\; dY_n \,]^T$$

be the vector of suggested movement of each landmark to match corresponding features of the image. The value of $d\mathbf{X}$ can be either based upon region or gradient information and normally is the normal distance to the closest image edge.

[9] In the following we shall use the notation of [Cootes & Taylor 92]

Finally, let us indicate by $T(s, \theta, \mathbf{x})$ the transformation that scales and rotates the model by s and θ, respectively, and by $\mathbf{X}_c = [X_c\ Y_c\ X_c\ Y_c\ \cdots\ X_c\ Y_c]^T$ the $2n$-element vector of equal landmark translations.

Now, since we have $\mathbf{X} = T(s, \theta, \mathbf{x}) + \mathbf{X}_c$, it is easy to show that by adjusting the landmark points by $d\mathbf{X}$ we have:

$$d\mathbf{x} = T((s(1+ds))^{-1}, -(\theta + d\theta), T(s, \theta, \mathbf{x}) + d\mathbf{X} - d\mathbf{X}_c) - \mathbf{x}$$

where the pose displacements ds, $\mathbf{X}_c$ and $d\theta$ are computed by simple geometric considerations from $d\mathbf{X}$ by averaging each landmark contribution, as detailed in [Cootes & Taylor 92].

These displacements are arbitrary (thus in general not consistent with the statistics of the shape model) and therefore $d\mathbf{x}$ is transformed into the space of the PDM modes of variations. From Equation (3.13) we have

$$\mathbf{x} + d\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}(\mathbf{b} + d\mathbf{b})$$

from which it follows that the model-consistent adjustment of the mode vector $\mathbf{b}$ due to $d\mathbf{X}$ is given by:

$$d\mathbf{b} = \mathbf{P}^T d\mathbf{x}. \tag{3.14}$$

In [Cootes & Taylor 92], it is pointed out that this method is equivalent to a least squares approximation of $d\mathbf{b}$.

All the procedure is iteratively repeated – until convergence is achieved – by updating position, scale and shape parameters by

$$\begin{aligned} X_c &\to X_c + w_t dX_c \\ Y_c &\to Y_c + w_t dY_c \\ \theta &\to w_\theta d\theta \\ s &\to w_s ds \\ \mathbf{b} &\to \mathbf{b} + \mathbf{w}_b d\mathbf{b}, \end{aligned}$$

where w_t,w_θ, w_s and $\mathbf{w}_b$ are weights that can be chosen to be either one, set to favour certain shape variation instead than others or for speeding up convergence by over-relaxation; further details can be found in [Cootes & Taylor 92]. In the experiments here the weights have all been set to one but in the method used in the next chapter, they are made to linearly decrease with the number of iterations to avoid instabilities (overshooting).

3.4.5 Fitting to unsorted set of points

In some cases, the data to which the PDM has to be fit, might present itself as a point data set, e.g. if the fitting has to be performed to *codons* as in the next chapter.

The fitting of the PDM to point data set is carried out in the same way as outlined in the previous section, the only difference being the way the $d\mathbf{X}$s are computed, notably by considering the difference between landmarks and corresponding data points.

However, there are some additional problems that have not been properly tackled in the literature on PDM (well, perhaps it was just not needed). The two main problems, exemplified in Figure 3.15, are:

- How to find the landmark-to-data correspondence necessary to evaluate the deviation for each point.
- How to cope with rank-deficient data, that is where a model landmark has no corresponding data point, a situation common when the data points do not span the whole object contour.

In fitting to raw images, it is the model itself that establishes correspondence between image features and landmark points. On the other hand, here the correspondence is more data-driven, since data points can be viewed as targets for each landmark.

For finding correspondence, a simple closest-point search has been employed with some shrewdness in order to avoid the same point being considered more that once by each model landmark. The complexity of this search is $O(MN)$ where N is the number of data points and M is the number of model landmarks; however, since M is constant

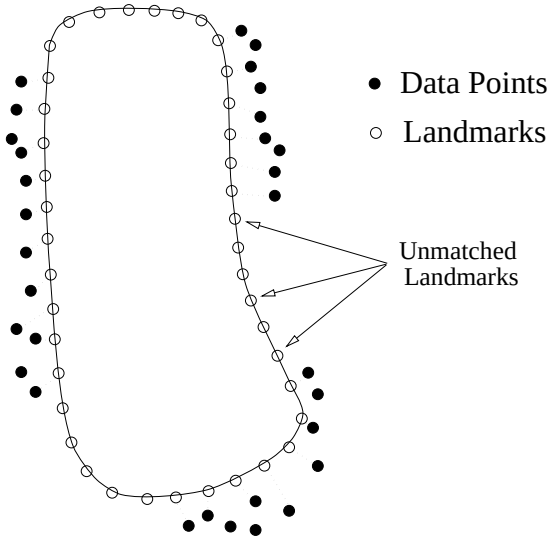


Figure 3.15: PDM fitting the point data set: the problem of correspondence. See text for details.

the search can be considered linear in the dimension of the data set. The details of this trivial search are irrelevant and therefore will not be discussed here.

A very intriguing possibility is to use the the elegant method proposed in [Scott & Longuet-Higgins 91] (reviewed in Appendix B), where a clever application of the singular value decomposition is used to minimise the overall sum of the squared landmark/datum distances, hence establishing sub-optimal correspondence; the method has also been succesfully used in [Shapiro & Brady 92] and [Sclaroff & Pentland 95]. Its application to our point-to-point correspondence problem for fitting PDMs seems a straightforward extension since both PDMs and the point-data set can be seen as two feature patterns to be matched. The main advantage is that it produces a *global* optimal mapping, as can be clearly seen in the two examples of Figure 3.16: in both cases, many landmarks (asterisks) do not pair up to the (geometrically) closest data points (circles) and yet the method manages to find a reasonable mapping. The use of such a method could also dramatically increase the robustness to poor initialisation. I have performed a number of experiments to test this possibility and the preliminary results look promising. However, it is not yet clear how the method performs when a different number of landmarks and data points is to be matched (also pointed out in [Shapiro & Brady 92]) and how can the

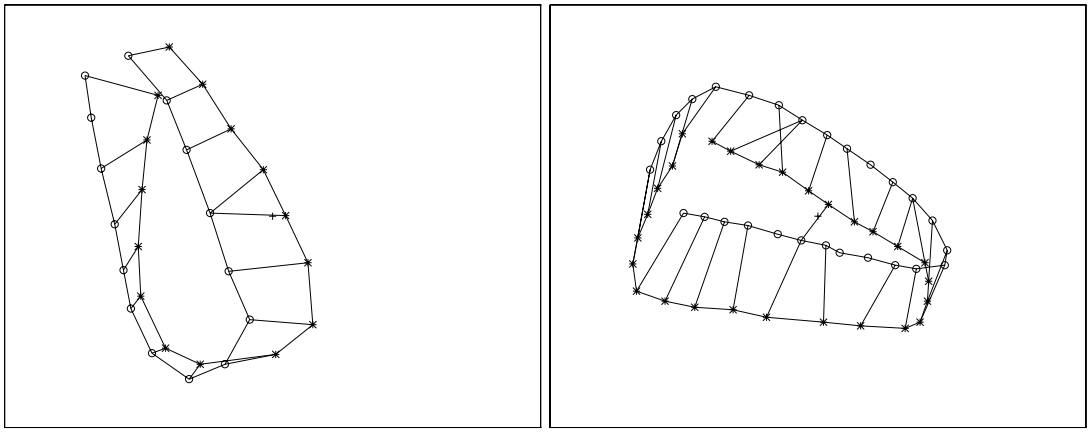


Figure 3.16: Correspondence problem. The use of the SVD method mapping. Many landmarks (asterisks) do not pair up to the geometrically closest data point (circles) and yet the method manages to find the correct mapping.

critical σ parameter that attenuates interaction between distant points be determined [Scott & Longuet-Higgins 91].

However implemented, this correspondence problem sometime slows down convergence due to occasional local flipping of mapping between neighbouring points and landmarks.

The other problem arises when some landmark has no correspondence in the data set. Referring to the standard fitting procedure outlined in Section 3.4.4, this inconvenient has been overcome as follows.

At each iteration of the fitting procedure, the landmarks-to-data correspondence is found and *only* the matching pairs are used to compute the $d\mathbf{X}$, the $d\mathbf{x}$ and finally the variation of the mode vector $\mathbf{b}$ as in Eqn. (3.14).

Formally speaking, if $\mathcal{I}$ is the set of k indices $\{i_1, j_1, i_2, j_2, \ldots, i_k, j_k\}$ to the vector of landmark coordinates $\mathbf{x}$ *having* correspondence in the data, and if we indicate by $\mathbf{A}_{\mathcal{I},*}$ the sub-matrix of a generic matrix or vector $\mathbf{A}$ obtained by assembling the k rows $i_1, j_1, i_2, j_2, \ldots, i_k, j_k$ of $\mathbf{A}$, the procedure outlined in Section 3.4.4 can still be applied if at each iteration we perform the following substitutions in the algorithm:

$$\mathbf{X} \leftarrow \mathbf{X}_{\mathcal{I},*}$$

$$\mathbf{x} \leftarrow \mathbf{x}_{\mathcal{I},*}$$

$$\mathbf{P} \leftarrow \mathbf{P}_{\mathcal{I},*}$$

By this simple method, it is possible to perform the fitting to both incomplete and complete data point sets. The satisfactory functioning of this procedure is assured by the dimensionality reduction performed by the principal component analysis that allows much of the structure of the shape to be recovered even when considerable parts of the contour are missing; models having this nice property are often termed as *self-symmetric*. This is a very desirable feature in object recognition and for perceptual grouping in particular, where we would like self-symmetries of object models to aid the recovery stage in presence of missing data.

3.4.6 Some fitting experiments

In this section only few illustrative examples are given, because in the next chapter PDM fits will be seen in the hundreds.

The initialisation is performed by fitting ellipses, as suggested in Section 3.4.3, to the point data set and a few preliminary iterations are used to better align model and data before performing the global iterative optimisation including the shape parameters. In the figures, the little lines show the correspondences between data points and landmarks.

In Figure 3.17, the data set has a bean-like shape; both sides are reasonably complete. The final result shows that the PDM well grasped the bending and the tapering of the shape.

Figure 3.18 show a similar case where a pronounced tapering is well recovered in few iterations, although the the final shape is slightly misoriented.

Figure 3.19 gives a very interesting case where the fitting returns a model that has nicely "completed" the missing part in the data in a perceptually plausible manner.

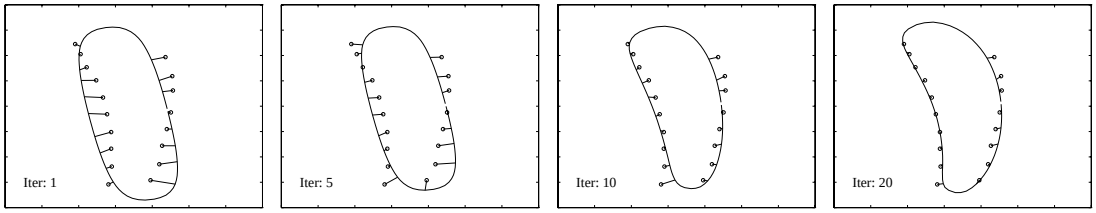


Figure 3.17: PDM fitting example. After few iterations, the bean-like point pattern is recovered and the missing part correctly "guessed" by the self-symmetries of the model.

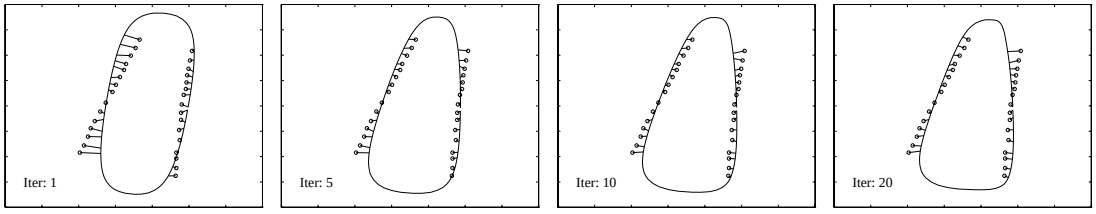


Figure 3.18: PDM fitting example. The tapering is properly recovered but an unexpected orientation displacement is still present after 20 iterations.

As we said before, this is due to the self-symmetric properties of this kind of models that help filling in gaps.

However, as shown in Figure 3.20, fitting is not always able to complete the shape in a perceptually acceptable manner: the right segment there is too small and noisy and consequently the final result is not a rectangle as desired. This problem would not be so pronounced (yet still present) if we fitted superellipses because of their more "rigid" structure.

In the literature, some of these problems have been tackled in various way, such as using a Bayesian integration of other image information [Haslam *et al.* 95], discounting sliding of control points or by employing model global contour information, as nicely summarised in [Cootes & Taylor 95]. In this work I have not directly experimented with these new techniques but it is evident that they are likely to further improve reliability.

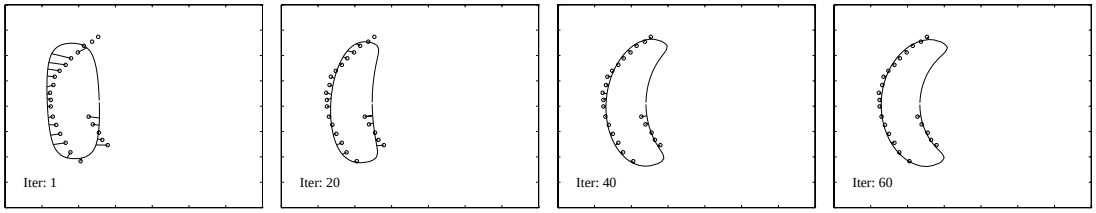


Figure 3.19: PDM fitting example. Under rather pronounced bending, the shape is correctly recovered but an higher number of iterations was required to achieve convergence.

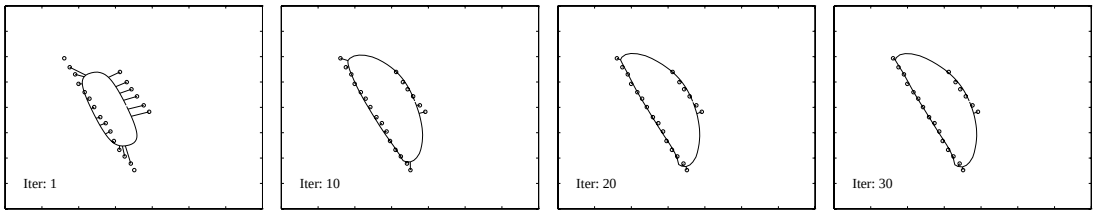


Figure 3.20: PDM fitting example. Here the shape is too incomplete and some point on the right-most PDM corners have been erroneously attracted to the longer segment, causing the result like the one shown in figure. In this case the model self symmetries have not been sufficient to extrapolate the correct shape.

3.4.7 Discussion and further work

The PDM built as in Section 3.4.2 allows to coarsely represent generic part contours in a reasonably general way. The model always keep an overall part-like structure that ensures that shape self-symmetries help shape recovery in case of occlusions and missing data.

We have seen that the PDM thus created has similar representational power to a DSE and its shape features are controlled by parameters with a precise geometrical meaning. The main concern was not to create a precise shape model for practically no objects can be exactly represented by DSEs and DSQs: a high degree of precision of representation is a lesser problem in generic shape analysis, which is the very domain DSE and DSQ are used for.

As pointed out in [Pilu *et al.* 96b], the spirit of the modelling method proposed is much more general; it suggests that, whenever convenient, complicated mathematical

shape models should be substituted with others with similar representational power to the foremost cause of fitting performance. For doing so, models such as PDMs can be used with a training set built with examples of the original model itself, as shown here in the case of superellipses. A related technique was presented in [Cootes & Taylor 94] but it that case PDMs were *sill trained on the data* in order to replace the clumsier Finite Element Models as soon as more examples were made available.

The extension of the method to superquadrics is straightforward. In the future, I also plan to apply the same model-trains-model strategy to other domains, for instance in the training on a shape class of high order polynomials (which can be fitted by closed-form least squares methods) in order to parametrise their dominant shape features.

3.5 Chapter summary

This chapter has addressed issues concerning three different models that can be used in a primitive-based part decomposition philosophy such as in [Pentland 90].

Section 3.2 presents what is perhaps the single most significant contribution of this thesis, that is the exposition and theoretical demonstration of the first ellipse-specific direct least squares fitting method, which was originally proposed in [Fitzgibbon & Fisher 95] as a curiosity. Direct least squares fitting of ellipses is a problem that was previously considered unsolvable. The novelty of the method is also confirmed by a most recent paper by [Rosin & West 95]), where non-linear iterative methods were considered as the only possible solution to the ellipse fitting problem. Many experiments are provided that show the relevance of the new method in comparison to other non ellipse-specific methods.

Section 3.3 illustrates the deformable superellipse model – a sub-case of the popular deformable superquadric model [Solina & Bajcsy 90], which allows a considerable increase in representational power with respect to ellipses. Deformable superellipses, to my best knowledge, have ever been used before in the vision literature. In this thesis, the superellipse model is used to generate a large synthetic set of part-like shapes in order to train a more general and versatile statistical part model.

Section 3.4 describes how a part-like statistical Point Distribution Model [Cootes *et al.* 91] is built by using randomly generated deformable superellipses as the training set. The idea of building such a model was dictated by the clumsiness of the deformable superellipse, which has been found hard to fit. At the beginning, this seemed hardly a contribution; later, however, the relevance was identified in having created what can be called a *linearly* deformable superellipse model, in the sense that the model simulates a superellipses but is, as any PDM, linear. The fitting method proposed in [Cootes & Taylor 92] has been modified to perform the fitting to unsorted and possibly incomplete point data sets. An interesting application of a recently developed technique for point-to-point correspondence is also proposed to improve the pairing of model landmarks and data points.

Chapter 4

Part-Based Grouping by Models

This chapter addresses the problem of instantiating part models from two-dimensional edge images. The proposed method is termed *part-based grouping* because it searches for part-consistent edge portions under the guidance of generic part models.

The input edge image is first partitioned into *codons* – contour portions of similar curvature [Hoffman & Richards 85]. Then, small *seed groups* of codons are used to initialise and pre-shape generic part models which are subsequently fitted to neighbouring codons; keeping an overall consistent part-like shape allows the determination of codon groupings consistent with the model. If a model is found not to have enough image support, it is rejected. This rejection rule, in the spirit of the Least Commitment Principle [Marr 82], is not severe and many grouping hypotheses are retained for a more sophisticated filtering stage which will the subject of the next chapter.

The organisation of the chapter is as follows. Firstly, the foundation of the method is given, followed by a review of part segmentation and grouping literature. Next, the method employed is outlined in Section 4.3, with a full description given in subsequent sections. The experimental section presents a number of grouping experiments for synthetic and real images. Finally, a discussion of the contributions, limitations and possible extensions to the method is given.

4.1 Foundations and historical background

Work on the decomposition of objects into their constituent parts can be tracked back to the early years of computer vision research in works such as by [Binford 71] and [Marr & Nishihara 78].

A considerable amount of this work hinged on representational issues, such as in [Marr & Nishihara 78] and [Nevatia & Binford 77] . Essentially, however, three theoretical works were to predominantly affect most of the subsequent research in object decomposition.

Marr [Marr 82] first stressed the importance of parts as an intermediate representation level which possesses invariance and stability properties.

Koenderink and van Doorn [Koenderink & vanDoorn 82], who were inspired by *renaissance* craftsman works and academic art, put forward the elegant concept that surfaces belonging to "distinguishable" entities of objects have elliptical properties, whereas the joints between them are of hyperbolic nature. In fact, they also make a strong case for *" [...] the hypothesis that vision grasps shape as a hierarchical structure of elliptic patches [...]"*.

In their fundamental work, [Hoffman & Richards 85], set out their simple yet important *transversality principle*, which states that *"when two arbitrarily shaped surfaces are made to interpenetrate they always end in a contour of concave discontinuity of their tangent planes"*. They also argue that *"part decomposition should precede part description"*, an idea destined to have important consequences in subsequent research. This philosophy was rather in contrast to primitive-based works in which part recognition and description was accomplished in a single indivisible stage: it suggests that solid parts can be segmented just by identifying high-curvature concave regions on the occluding contour of an object.

Unfortunately, works following Hoffman and Richard's idea have invariably used either closed outlines or silhouettes of objects and operated more or less locally by finding *convex dominant points* (CDP) from which part presence had to be inferred. But, as is well known by the computer vision community, in most but the simplest cases, the

extraction of silhouettes is virtually impossible, thus rendering most proposed computational methods prone to early criticisms on the unrealistic assumptions made on the input data. Curiously, the *objectness*, i.e. enclosure of material, that is physically associated to parts has often been explicitly neglected in this kind of computational approach, but nonetheless unconsciously adopted: silhouettes in fact do bound the material composing the object!

Another category of approaches relies on *perceptual grouping*, mainly due to Witkin and Tennenbaum [Witkin & Tenenbaum 85] and Binford [Binford 71]. Perceptual grouping is the procedure of *"finding and grouping salient non-accidental features that would occur with non-zero probability if produced by a single object or process"* [Lowe 85] . Perceptual grouping in computer vision stems from the work in Gestalt perceptual organisation [Kohler 59], which strongly advocates that such a process occurs pre-attentively in biological vision systems. In fact, perceptual grouping not only has been relatively successful in explaining some well known Gestalt phenomena [Gibson 79, Kohler 59], but impressive results have also been produced in vision systems such as [Mohan & Nevatia 92], [Jacobs 96] and [Lowe 85]. Actually, one of the main advantages of perceptual grouping is that it does not require infeasible assumptions on the input image and the kind of objects present in it.

As far as part segmentation goes, however, classical perceptual grouping approaches have shown some limitations. In particular, the inability to cope with bent parts in the case of convex grouping, or to perform boundary completion when a substantial portion of part contour is missing. The latter problem is observable in many works [Mohan & Nevatia 92, Rao & Nevatia 89, Stein & Medioni 92], where the two whole sides of parts have always to be available. To overcome this problem, some heuristic criteria have often been used, such as the "U" completion rule [Stein & Medioni 92, Mohan & Nevatia 92].

One common aspect of these perceptual grouping approaches is that the objectness is not explicitly represented; here it is argued that the "thing-like" nature of parts can be better grasped by using "thing-like" models as a guide to all processing stages.

A conjecture made by many researchers in perception such as in [Biederman 87] and [Marr & Nishihara 78], which is also the foundation of part-based recognition, is that

complicated shapes are grouped and perceived by decomposition into simpler shapes. Although no claim of biological plausibility is made here about the use of simple models, it is a rather striking fact that most pre-attentive global perceptual grouping phenomena occur for simple forms. The famous Kanisza experiments, for instance, invariably use simple shapes like triangles, circle and spirals.

The part-based grouping method presented in this thesis proposes a new computational approach in which this conjecture is followed by using simple geometrical models of what these shapes are expected to be. Fundamentally, the method fits in the Hoffman and Richards' decompose-before-describing philosophy, but it solves the grouping problem by blending it with the use of simple 2D primitives, clearly more in the spirit of Koenderink and Van Doorn's ideas.

No restrictions are placed on the input data: the method can inherently cope with ordinary edge images with a reasonable amount of cluttering, fragmentation and interior edges.

After a brief review of past work in both part-segmentation and perceptual grouping, the approach is first overviewed in Section 4.3 and detailed in later sections.

4.2 Previous related work

In this section a brief review of part decomposition and perceptual grouping literature is given. The literature regarding these two problems is rather vast and therefore I shall concentrate on works that are closely related to this thesis, namely the decompose-before-describing strategy for part segmentation and medium-level feature grouping.

4.2.1 Part decomposition literature review

Previous works in decompose-before-describing framework for part recognition are categorised and briefly reviewed in the following.

Symmetry axes or skeleton techniques first locate symmetry axes in various ways and then associate each axis of symmetry to an object part. Such methods were investigated in, e.g. [Blum & Nagel 78], [Burns *et al.* 94], [Pizer *et al.* 94], and

elegant computational methods were given, notably, in [Rom & Medioni 93] and [Zerroug & Nevatia 94]. Rom *et al.* iteratively decompose parts starting from short axes (supposed to correspond to details) all the way up to the main symmetry axis. Zerroug *et al.* use instead a method that initially makes use of symmetry axes but performs actual part decomposition by looking for "L" and "T" junctions in the edge image, a technique which cannot be reliably used for real images. However, the symmetry axis, although originating from two (skewed-)parallel curves likely to belong to a single part, does not *per se* explicitly embed objectness.

To my best knowledge, no method has been devised as yet to *reliably* extract symmetry axes for objects described by other than their silhouette (or silhouette contour); in fact most research, perhaps with the exception of [Mohan & Nevatia 92], limit themselves to the recovery of the symmetry axes of neat objects and do not deal with the selection of multiple symmetries or incomplete axis in a reasonable manner, if this is done at all.

Contour-based techniques normally detect, in the spirit of the transversality principle, convex dominant points (CDP) in various ways and then infer parts by putting CDP in proper correspondence. Such an approach has been followed in, e.g., [Freeman 78], [Bennamoun & Boashash 94] and [Siddiqi & Kimia 95]. Although intuitive and yielding reasonable results, these methods work only with silhouette input.

Primitive-based techniques centre on detecting "objectness" in the data by using compact, generally convex, simple models such as ellipses. The enormously popular work by [Pentland 90] used a simple template matching strategy to generate hypotheses which were later filtered out to leave the most significant ones. Despite the trivial generation of hypotheses and the use of silhouettes, his work inspired many researchers; for instance, in [Hara *et al.* 92] a simple growing strategy is used to extract elliptical patches from silhouettes. Thus far, however, all the primitive-based techniques have used just silhouettes as input.
As we shall see later, the method presented in this thesis bears considerable resemblance to the spirit of this primitive-based approach.

Scale-space approaches, nicely formalised by [Lindeberg 94], have also been proposed for part detection but unfortunately the results are so far not very promising. In fact,

it is not clear how the multi-scale smoothing process can possibly elicit part-like blobs for complicated object appearances.

Graph-theoretic methods, notably [Shapiro & Haralick 79], exploit various clustering techniques of nodes representing points. These methods have been explored only for closed contours. It would be a research topic on its own right to investigate their use for cluttered and incomplete edge images but, to my best knowledge, nobody has yet ventured along that avenue.

Finally, [Kimia *et al.* 95] proposed, perhaps after Leyton's research on causal theory of shape [Leyton 92], a diffusion method that would naturally perform part segmentation. Needless say that such a paradigm relies upon a silhouette as input.

4.2.2 Perceptual grouping

Although perceptual grouping (PG) for recognition had been acknowledged as extremely important [Jacobs 96], most works have dealt with very low-level point or edgel grouping, such as in [Sha'ashua & Ullman 88], line completion [Cox *et al.* 92], dot grouping [Link & Zucker 88], and in general all Gestalt properties. However, low-level grouping, has perhaps not yielded as much increase in performance to vision systems as was originally expected.

Lowe was probably one of the first vision researchers to explicitly incorporate simple perceptual grouping rules to increase speed and accuracy in his SCERPO vision system [Lowe 85]. However, chiefly simple local properties (with the exception of parallelism) were used, the main purpose of using PG in his system being the reduction of the search complexity in the database of objects.

Some good attempts have been made in integrating more global information into the grouping process, such as by recovering *convex groups* [Jacobs 96, Huttenlocher & Wayner 92] or using symmetry (e.g. in [Mohan & Nevatia 92]).

Jacobs identifies convex groups of linear edge segments likely to belong to a single object by accounting for the percentage of covering of their convex hull [Jacobs 96]. He estimates that such a grouping would speed-up traditional object recognition systems by a factor of 100 to 1500. Huttenlocher solves the same problem by using local

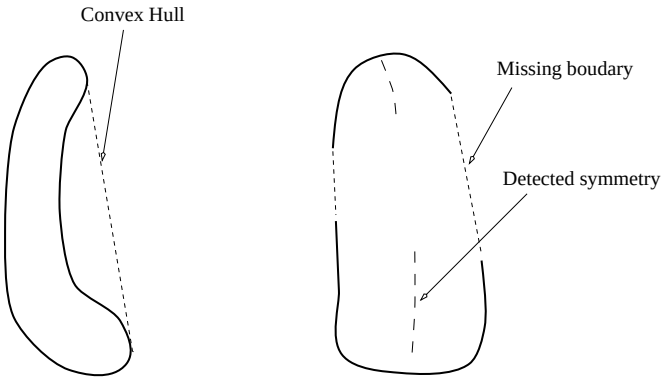


Figure 4.1: Inadequacy of convex hull (left) and symmetry (right) for the grouping of bent and convex parts, respectively. The convex hull cannot describe bent parts whereas symmetry has problems in dealing with occlusions.

relationships between adjacent edges [Huttenlocher & Wayner 92] . In his method, the $O(n \log n)$ complexity is achieved by making purely local decisions about how to extend a convex group; the inclusion of more global properties was left to future work. It has to be noticed that, although effective, convex grouping cannot deal with bent shapes, for which grouping should work as well as for strictly convex parts, as shown in Fig. 4.1-left.

Symmetry-based grouping methods, such as [Stein & Medioni 92], [Zerroug & Nevatia 94], [Rao & Nevatia 89] and [Mohan & Nevatia 92], have become quite a popular method of hypothesising objects in scenes. As known from Gestalt psychology, symmetry is a non-accidental property that carries significant statistical information and often arises from an object in the scene. Possibly, the best work in perceptual grouping based on symmetry is that by [Mohan & Nevatia 92], in which a hierarchical vision system was built that segments a scene into generic objects. The results they presented yet again clearly show how the use of perceptual organisation allows vision systems to cope with unknown objects and cut down complexity. However impressive, as a result of the symmetry-based approach their method and other similar ones cannot properly cope with heavy fragmentation or occlusion, as shown in Fig. 4.1-right.

Amongst other things, the aim of this thesis was to investigate how more complicated global information could be included in the grouping stage; as we shall see in the next

```
Partition image contour into codons (Sec. 4.4)
Find small part-plausible seed groups of codons (Sec. 4.5)
for each seed group do
      Initialise the part model to the seed group  (Sec. 4.6.1)
      Pre-shape the part model to the seed group  (Sec. 4.6.2)
      Find supporting codons to the pre-shaped model (Sec. 4.6.3)
      Fit the part model to the additional support (Sec. 4.6.4)
end for
A set of part hypothesis is now available
```

Figure 4.2: Pseudo-code of the model-guided part-based grouping method proposed in this thesis. See text for details.

section, this has been achieved by using generic shape models to guide the grouping of features.

4.3 Rationale of the approach

This section outlines the rationale of the approach that is going to be detailed in the next sections; Fig. 4.2 gives the *pseudo-code* of the algorithm.

Representing objectness by geometric models

Let us take a look at Figure 4.3-left. To the eyes of the human, the tree would be grossly described by two parts: the trunk and the foliage. Clearly this part subdivision is independent from what Biederman called *structural details* [Biederman 87]. To achieve this abstract part-level description, a computer system should not only employ some means for "smoothing" the shape but also have a notion of the essential "thing-like" nature of parts [Pentland 90]. This is also valid for the three-dimensional case as, according to Huffman and Richards's theory of parts[1] [Hoffman & Richards 85], solid parts can be inferred from their 2D projection by looking for non-accidental invariant properties in edge images.

The "thing-like" nature of parts, also called *objectness*, had often been neglected as a guideline to the computational study of the part segmentation problem until the work

[1] Also supported by practical evidence.

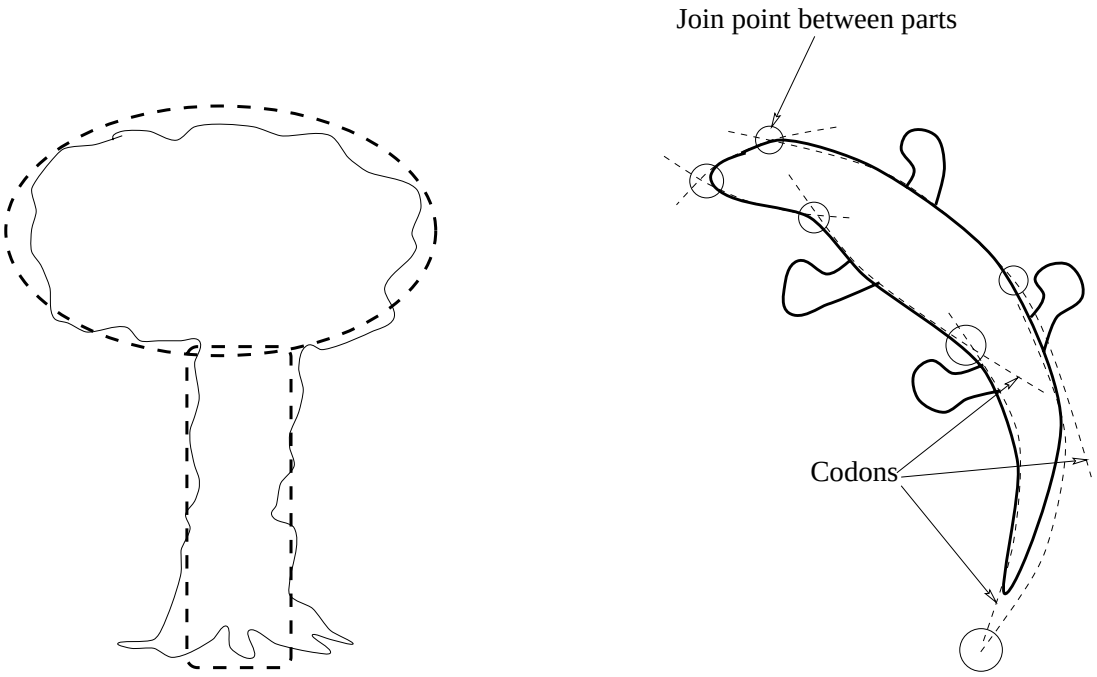


Figure 4.3: Left: Natural part decomposition; to the human eye the tree is grossly described by trunk and the foliage, independently from structural detail. Right: Codon extraction and part decomposition (lizard example after [Subirana-Vilanova & Richards 91]); neglecting the paws, the dashed lines are curves whose intersections naturally identify part joints.

of [Biederman 87] and, in particular, [Pentland 90], who argued that objectness can also be expressed by *a set of generically applicable part models* and this line of thought is the hinge of this thesis.

Objectness is represented here by the closed contour of the simple generic part PDM presented in 3.4.

Explicit use of codon properties

However, as [Pentland 90] put it, there is no known computational model to *"begin immediately with recognition of part models"*.

The computational infeasibility of a method that directly look for parts in an image suggests that perhaps it is necessary to step one level back from whole-part models in an hypothetical representational hierarchy of objects.

The most natural level is the one of *codons* which are medium-level curve primitives that are *"pieces of boundary bounded by negative curvature minima"* that delimit parts naturally and in compliance to the transversality principle [Hoffman & Richards 85].

Figure 4.3-right shows an example of codons for the case of a silhouette of a lizard (after [Subirana-Vilanova & Richards 91]); neglecting the paws, the dashed lines are curves that represent codons and the small circles are the intersections between them. It can be seen that the intersections naturally identify part joints.

As said in Section 4.2, most part segmentation approaches somehow by-pass an explicit detection of significant codons by leaping directly to the determination of convex dominant points, which are then used to infer part decomposition. But convex dominant points have a local nature and do not express relationships to other points, which is a vital component of part segmentation. As an example, if all lines were taken off the lizard in Figure 4.3-right, it would be impossible to tell the shape only from the position of the convex dominant points (circles).

In this respect, codons turn out to be very important: they are rather global in nature, are easily recoverable and, by their sheer definition, *bound single parts*, that is a codon cannot belong to two separate parts except for few accidental cases. With a certain degree of confidence, it can be said that codons are the golden choice as an intermediate representation level for generic part recognition.

In this thesis, an approach is indeed followed that explicitly recovers and represents codons by second order polynomials. Although more sophisticated multi-scale (or *appropriate* scale) methods could have been used for codon extraction, a simple single-scale method has been used that provides an acceptable trade-off between simplicity and performance.

Pre-grouping of codons

Once codons are available, a method should be devised for "grouping" codons belonging to single parts.

The use of a codon-like contour partitioning scheme is certainly not new, having been

widely used in various forms for these kinds of problems. However, most works – notably symmetry-based – assume that each codon covers most of the part sides. Unfortunately, this is not the case in real images: often codons are over-segmented, whole boundary segments missing, and marking, shadow and shading edges are always present.

A specific aim of this research was to devise a technique that performs part-based completion of missing contour portions and is able to cope with cluttered and fragmented images. The strategy followed here is to use generic part models to guide the search for codon groups that are likely to represent actual parts, henceforth called *part-plausible groups*.

In order to recover these part-plausible groups, a simple-minded approach would be a brute-force generation of all possible combinations of codons and ranking them according to an "object-ness" measure. Of course this is not a feasible solution because of its exponential complexity and therefore another method had to be devised.

Codons can be considered as *seeds of perception* [Brady 87] from which more and more complicated descriptions of the images are constructed. Our final aim is to achieve a level where there are groups of codons associated with each part in the image; in an intermediate stage, small groups of codons, termed *seed groups*, can be recovered that are likely to give significant information on the part-structure. This stage is called here *pre-grouping*, because it is a prelude to the real, more global grouping that in this thesis is performed under the guidance of generic part-models, as explained next.

Part pre-shaping and fitting

As just said, once seed groups of codons are available, models are first pre-shaped and then fitted to the image data.

First, coarse positions and orientations of the part-like models are determined by fitting ellipses to all the pixels belonging to seed groups as explained in Sec.3.4.3.
Successively, the PDMs are *pre-shaped* to the seed group of codons; in this phase coarse bending and/or tapering estimates are recovered along with positions and dimensions. Then codons are recovered that are somehow in agreement with the pre-shaped model

instance and finally a global fitting is performed that deforms the shape to conform to all the image evidence. Section 4.6 will discuss these matters in more detail. Pre-shaping can be seen as a way of reducing complexity and facilitating convergence, much as has been done in, e.g., hand pre-shaping for robot grasping [Wren 96].

Many hypotheses are created but the great majority of them will represent the contour data poorly due to the lack of image evidence and can be discarded straight away. However, a number of good or plausible hypotheses end up contending for describing the image evidence; the filtering of these hypotheses to produce part segmentation is the subject of the next chapter.

The outcome of this procedure is to effectively produce a part-based grouping. It is necessary to stress that this model-guided grouping method is complementary to other grouping techniques, in the sense that it cannot alone solve the grouping problem. These matters will be discussed more in Section 4.8.

4.4 Codon extraction

The importance of codons in the context of part recognition was discussed in Section 4.3. Codons are defined as being pieces of contour bounded by negative curvature minima [Hoffman & Richards 85] but this plain definition is too fuzzy to be operative, probably because it originated from the silhouette-input frame of mind, where negative and positive signs of curvature can be unambiguously determined.

Although it would have been possible to draw from the wealth of techniques available in the literature for contour salient partitioning (e.g. [Lowe 88], [Teh & Chin 92], [Saint-Marc & Medioni 88], [Fischler & R.C.Bolles 86], [Rosin & West 95]), in this thesis I have used a variation of the simple but efficient iterative-line-endpoint-fit-and-split method [Ramer 72].

The following subsections describe how codon are represented, the partitioning method and some examples and comments.

4.4.1 Codon representation

A representation of codons by second order polynomial curves was chosen. The reasons for this choice are:

- Second order polynomials are good qualitative piecewise approximations of object contours [Rosin & West 95];
- They keep the sign of curvature, as the codon definition demands;
- They have a smoothly varying curvature and therefore, as suggested by the transversality principle [Hoffman & Richards 85], boundaries of different parts should be described by distinct polynomials;
- The recovery procedure is fast and can normally smooth out small details; therefore, only a limited amount of preprocessing, if any, to the raw edge images is needed.

In the following two subsections the actual algorithm is described along with some experiments.

4.4.2 Iterative polynomial endpoint fit and split

Following an edge detection and tracking phase, codons are extracted by an extension to second order polynomials of the *iterative line endpoint fit and split* (ILEFS) method that was first presented in [Ramer 72] (used also in [Lowe 85]) that I have called *iterative polynomial endpoint fit and split*. A recent paper [Rosin & West 95] advocated against the use of the ILEFS method for it sometime produces counter-intuitive results but, although I agree with their remark, it should be noticed that their context was different in that a precise segmentation into curve primitives was sought; here this is not at all important, because polynomials are used to approximate image contours rather that constituting *features* to be matched.

The original algorithm was designed to perform a polygonal approximation of a planar curve by recursively splitting, at points of maximum deviation, lines that link the curve's endpoints; for its termination, it requires only a maximum distance threshold,

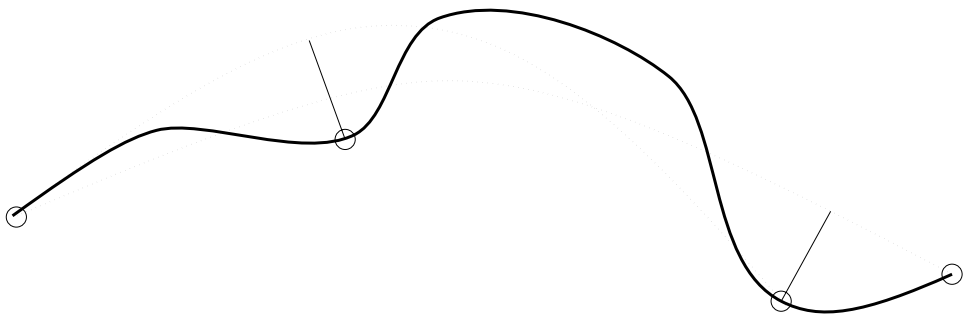


Figure 4.4: Example of the first two iterations of the iterative polynomial end point fit and split algorithm. The least squares polynomial passing through the two endpoints is recursively split at points of maximum deviation. See text for details.

indicated by d_{max}. Here, I employ the same procedure but instead of using straight lines the endpoints are linked by a second order polynomial passing through the endpoints that best approximates, in the least square sense, the whole curve. The first two iterations of the method are exemplified in Figure 4.4.

Codons are represented by second order parametric polynomial curves but the method can be easily extended to higher order curves, provided the inclusion of a constant-sign-of-curvature constraint in the least squares fitting procedure.

Let an edge be represented as a sorted sequence of N points $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ and the codon $\mathcal{C}_k$ passing through the points (x_1, y_1) and (x_N, y_N) as the second order parametric polynomial

$$\mathcal{C}_k = \begin{cases} x = a_x t^2 + b_x t + c_x \\ y = a_y t^2 + b_y t + c_y \end{cases} \qquad \text{with} \quad t = 1 \ldots N. \tag{4.1}$$

It is well-known that that such a polynomial is a parabola [Sederberg *et al.* 84, Rosin & West 95]. The coefficients are given by $\{a_x, b_x, c_x\} = PF(\mathbf{t}, \mathbf{x})$ and $\{a_y, b_y, c_y\} = PF(\mathbf{t}, \mathbf{y})$, where $\mathbf{x} = [x_1 \cdots x_N]^T$, $\mathbf{y} = [y_1 \cdots y_N]^T$, $\mathbf{t} = [1 \cdots N]^T$ and PF is the polynomial fitting function detailed in Appendix D.

For any point (x_i, y_i) let

$$d_i = \sqrt{(x_i - a_x i^2 + b_x i + c_x)^2 + (y_i - a_y i^2 + b_y i + c_y)^2}$$

be its distance to the approximating polynomial and let m be the index for which

$d_m = \max_i\{d_i\}$. If $d_m > d_{max}$ then the sequence of points $\mathcal{S}$ is split into two subsequences $\mathcal{S}_1 = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ and $\mathcal{S}_2 = \{(x_{m+1}, y_{m+1}), \ldots, (x_N, y_N)\}$ and to these two sequences the procedure is recursively applied until each subsequence has a maximum error below d_{max}.

4.4.3 A few experiments and comments

Here, a few examples are presented and the use of more sophisticated methods for achieving codon extraction are suggested.

Figure 4.5 shows the codon extraction for three different values of d_{max} for three real images – one in each column – of a handset, a hand and a multi-object image (a wooden stick, a marker and a screw-driver). The values of d_{max} are expressed in image pixel units. It can be noticed that the overall structure is kept for large changes in d_{max}. In all the experiments of this chapter values of $d_{max} = 2$ for 128x128 images and $d_{max} = 4$ for 256x256 images have been used.

Probably the state of the art in single-scale contour extraction is proposed in the recent work by [Rosin & West 95], where image contours are partitioned into a combination of representations such as line, polynomial, elliptic and superelliptic; such a technique could be used to significantly improve codon extraction. If objects or parts at different scales were in the image, a multi-scale (or *appropriate* scale) strategy [Saint-Marc & Medioni 88] should be employed in order to ensure that the partitioning will not be either too fragmented or too coarse. Of course, a tangent continuity constraint at the joints could be easily introduced, although it should not improve performance significantly. Moreover, further processing stages will be greatly advantaged by the use of contour completion techniques (such as [Cox *et al.* 92] and [Sha'ashua & Ullman 88]) prior to this codon extraction phase.

Since the part grouping method proposed here has shown considerable resilience to the quality of the codon partitioning of the edge image, the study or implementation of a more refined technique is left for future work.

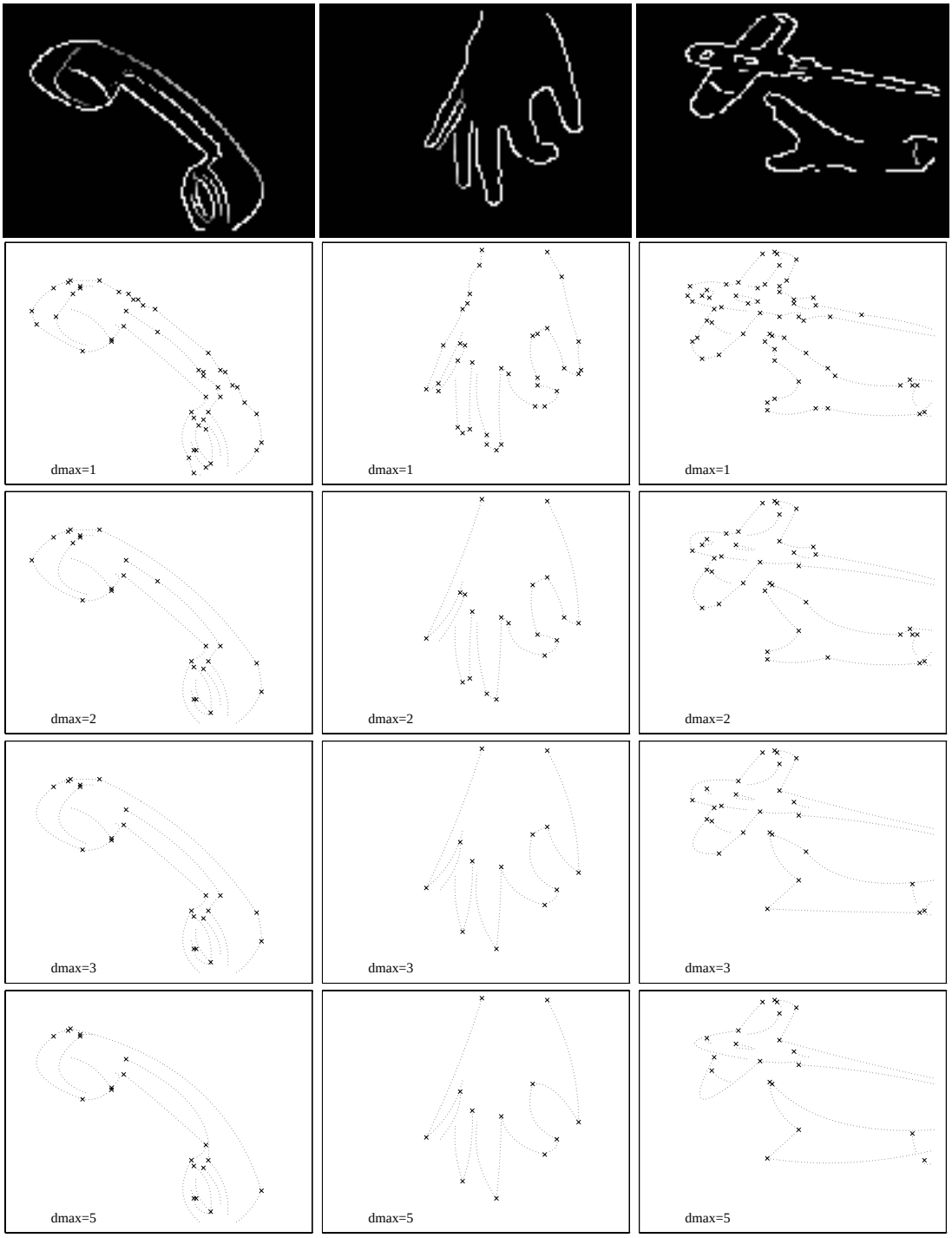


Figure 4.5: Codon extraction: Experiments with different d_{max} for three real images (one in each column) of a handset, a hand and a multi-object image (a wooden stick, a marker and a screw-driver). The values of d_{max} are expressed in image pixel units. It can be noticed that the overall structure is kept for large changes in d_{max}.

4.5 Codon pre-grouping

Finding codons bounding a single part is the aim of part-based grouping. For achieving this, in the proposed model-guided grouping strategy, part-like models have to be somehow fitted to the right codon data. As outlined in Section 4.3, the critical phase of model initialisation and pre-shaping is achieved by fitting PDMs to small seed groups of codons.

On one hand, one could create all possible 2^N combinations of N codons[2] and try to fit models to them. Although this would make sure that all the right hypotheses are generated, it is not feasible not even for objects with a very small number of codons. For instance, if 10 codons were present, the total number of hypotheses thus generated would be 1024. On the other hand, it might be that only one codon gives enough information for a part hypothesis to be created; however, this would be effective only for few cases, for example for a finger represented by a single codon, provided that a perfect codon extraction of the edge image is achieved.

More sensibly, groups of codons are better chosen according to heuristic criteria. We can call this phase *pre-grouping* since it involves finding part-plausible configurations of small number codons, possibly subsets of larger correct grouping hypotheses.

The fundamental postulate that justifies this pre-grouping strategy is: *a properly chosen small number of codons gives enough structural information for simple part shapes to be recovered*. This assumption is supported by practical and experimental evidence, which is extensively given in Section 4.7.

This pre-grouping technique is used in other contexts, such as circle or ellipse fitting (e.g. [Rosin & West 95]), where small contour portions are used to initially fit the model and then new evidence is sought by looking for other edges matching (more or less closely) the fitted model. In the case of deformable model fitting, this technique is starting to gain ground especially for complicated deformable models, such as for face recognition and medical image analysis. However, the features used to initialise the deformable models tend to be highly informative and often are in stable relative

[2] The power set of the set of codons.

position with respect to each other, which is somewhat opposite what happens here, where codons are very simple to extract but have a poor information content. In [Cootes & Taylor 96], an excellent short review of some of these recent techniques is given along with a new method.

Coming back to our problem, currently the seed groups simply consist of *all the possible pairs* of the $N' \leq N$ codons whose length (in pixels) is greater than 10% of the image size; these pairs are called *seed pairs*. Of course, by making this choice a relatively large number ($\frac{N'^2}{2}$) of of seed pairs (and therefore hypotheses) are produced and most of which will be either meaningless (i.e. not corresponding to an actual part) or duplicates. However, if enough structural information is available *in the edge image*, good groupings will always be produced that allow the next pre-shaping and fitting stages to recover the parts in the image and the experiments of Section 4.7 confirm this assertion well.

I have investigated some heuristics for reducing the number of generated seed pairs, since a smaller number of them would mean less PDMs to fit and faster filtering (see next chapter). These heuristics aim at not including pairs of codons that are very likely either *i)* not to belong to an actual part or *ii)* not to give enough information about its structure. Although a preliminary study reveals that the reduction in the number of generated pairs is substantial, the heuristics have not actually been robustly implemented; if they were, they would affect just time performance, which is anyway not a key issue at the present stage of the research.

The simple strategy of using seed pairs has obviously some limitations. If the edge image is over-partitioned into codons, pairs might not give sufficient structural information for the PDM to be pre-shaped[3]. This problem would manifest itself especially when parts are too bent because, if the pre-shaping is wrong, other additional information needed to carry out the final fitting might not be found. However, the idea could be greatly extended by forming seed groups using *early stages* of convex grouping methods (such as [Jacobs 96] or [Huttenlocher & Wayner 92]) and then letting the part models do the more expensive job of imposing global consistency as described in this thesis. This extremely interesting avenue is left for future work.

[3] Model pre-shaping is in fact performed on the seed groups; see next section.

4.6 Model fitting

This section discusses how the generic 2D part model is first initialised and then pre-shaped to the seed groups of codons, how further image evidence is found and finally how the part model is finally fitted to produce a grouping hypothesis.

The 2D part model chosen for this stage – the statistical PDMs of Section 3.4 – was suggested by these simple considerations:

- They have a "part-like" nature and structure;
- They constitute a good first order qualitative approximation to the outer contour of a large class of parts;
- They have inherent self-symmetry properties that help fill-in gaps and deals with severely incomplete boundaries;
- They are easy and efficient to fit.

As shown in Section 4.3, the generic part PDMs we use cannot represent complicated shapes such as the one in Fig. 4.6. Although it is true that the training set used could have been enriched with more complicated examples, I wish to stress that, as we shall see in the next chapter, in this framework it is more important how a hypothesis compares to others, rather than the score it achieves in isolation. Furthermore, complicated models would be more difficult to fit than those currently used, especially in the pre-shaping phase.

The fitting of the 2D part models has been extensively discussed in Chapter 3, but a few more specific notes will be added in the following.

4.6.1 Initialisation

Once the seed groups of codons are available, in order to perform model pre-shaping it is first necessary to have a coarse estimation of the model position, orientation and dimension. This initialisation need not be extremely precise but large errors might

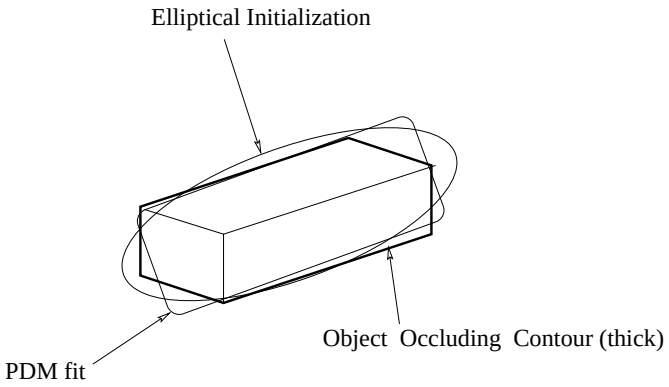


Figure 4.6: Ellipse and generic part PDM fitting to the outer contour of a block. None of them can account for the pronounced end effects but clearly the PDM can better represent the two sides.

jeopardise, if not impede, convergence of the iterative fitting stages soon to be described.

Through extensive experimentation, it has been confirmed that the initialisation by ellipse fitting to all the pixels of the seed groups of codons – as described in Section 3.2 – is a much better solution than the ordinary centroid method since, because of the nature of the codon data, it is essential that the initialisation is able to "guess" what the model should look like from incomplete data. The centroid method would cope well with blob-like data but not with incomplete edges. Moreover, initialisation by ellipse fitting also offers the advantage of giving, by itself, good shape information on a large variety of parts of natural objects. As a matter of fact, in early stages of the research and before the introduction of PDMs, parts were all approximated by ellipses and good results were also obtained in the filtering stage that is discussed in the next chapter.

However, as commented in Section 3.2, there is certainly much space for improvement here: the initialisation by ellipses does have some drawbacks, especially in the case of convergent codons (such as in Fig. 3.14).

Since it seems hard, if not impossible, to quantitatively ascertain the performance of this stage, a large number of initialisation examples are given from Figure 4.7 through Figure 4.10 ; the results are discussed in the captions.

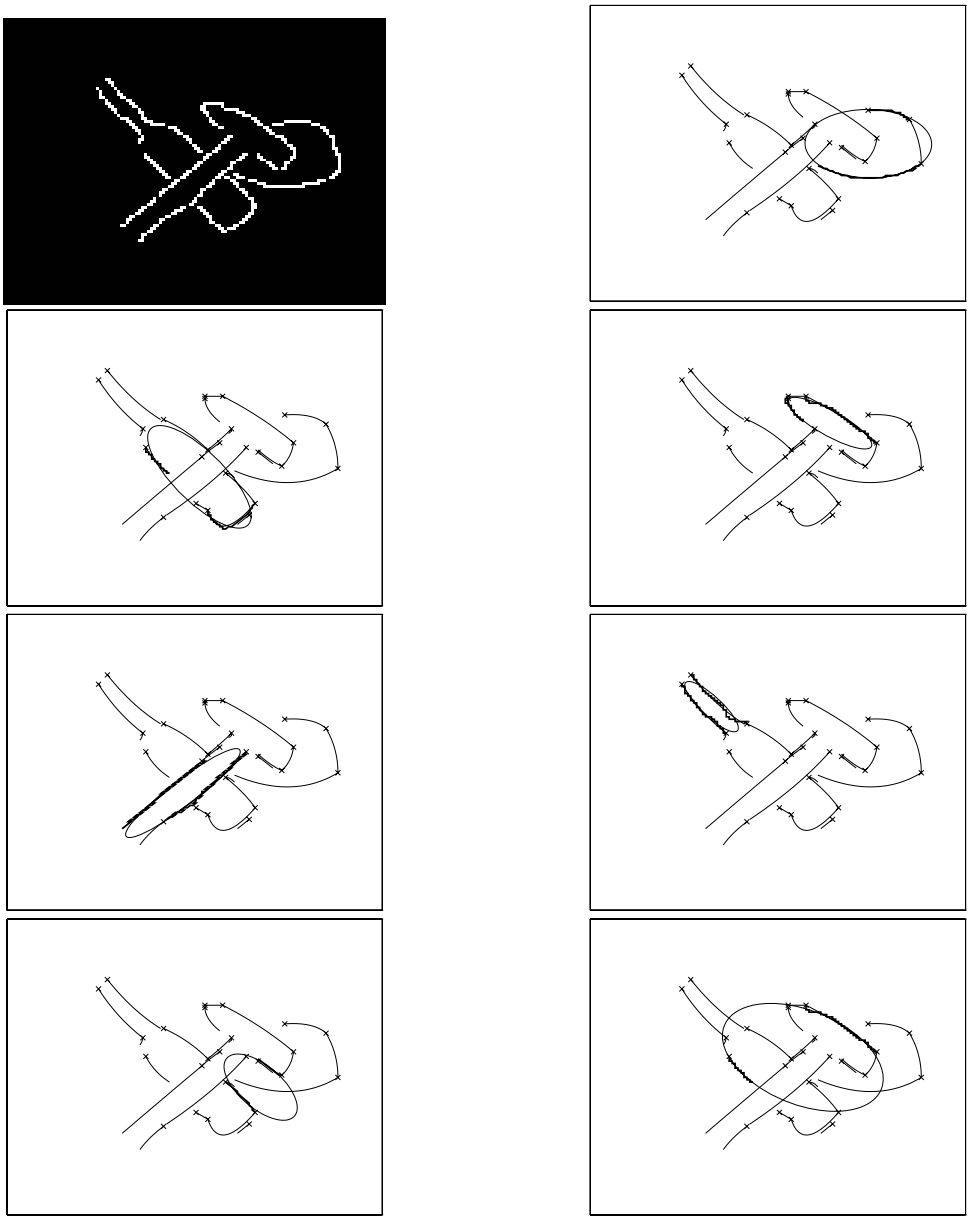


Figure 4.7: Seven initialisation examples to pairs of codons for the bottle and hammer example. The pairs of codons are represented by the rugged lines. It can be noticed that good ellipse initialisations are produced for the actual parts.

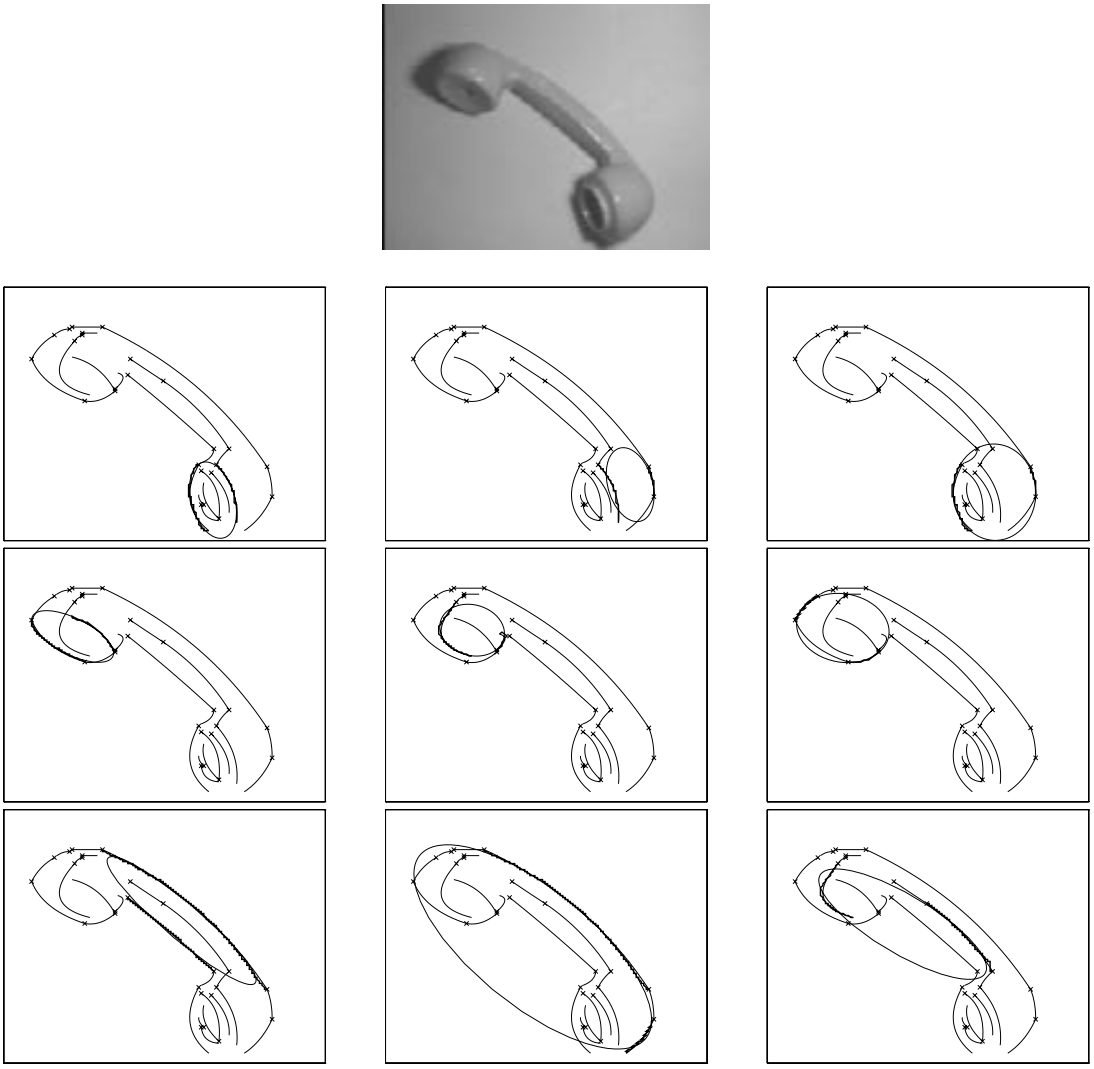


Figure 4.8: Nine initialisation examples to pairs of codons for handset . The pairs of codons are represented by the rugged lines. Good ellipse initialisation are produced for the actual parts, that are the handle, and the quasi-circular mouth and ear pieces.

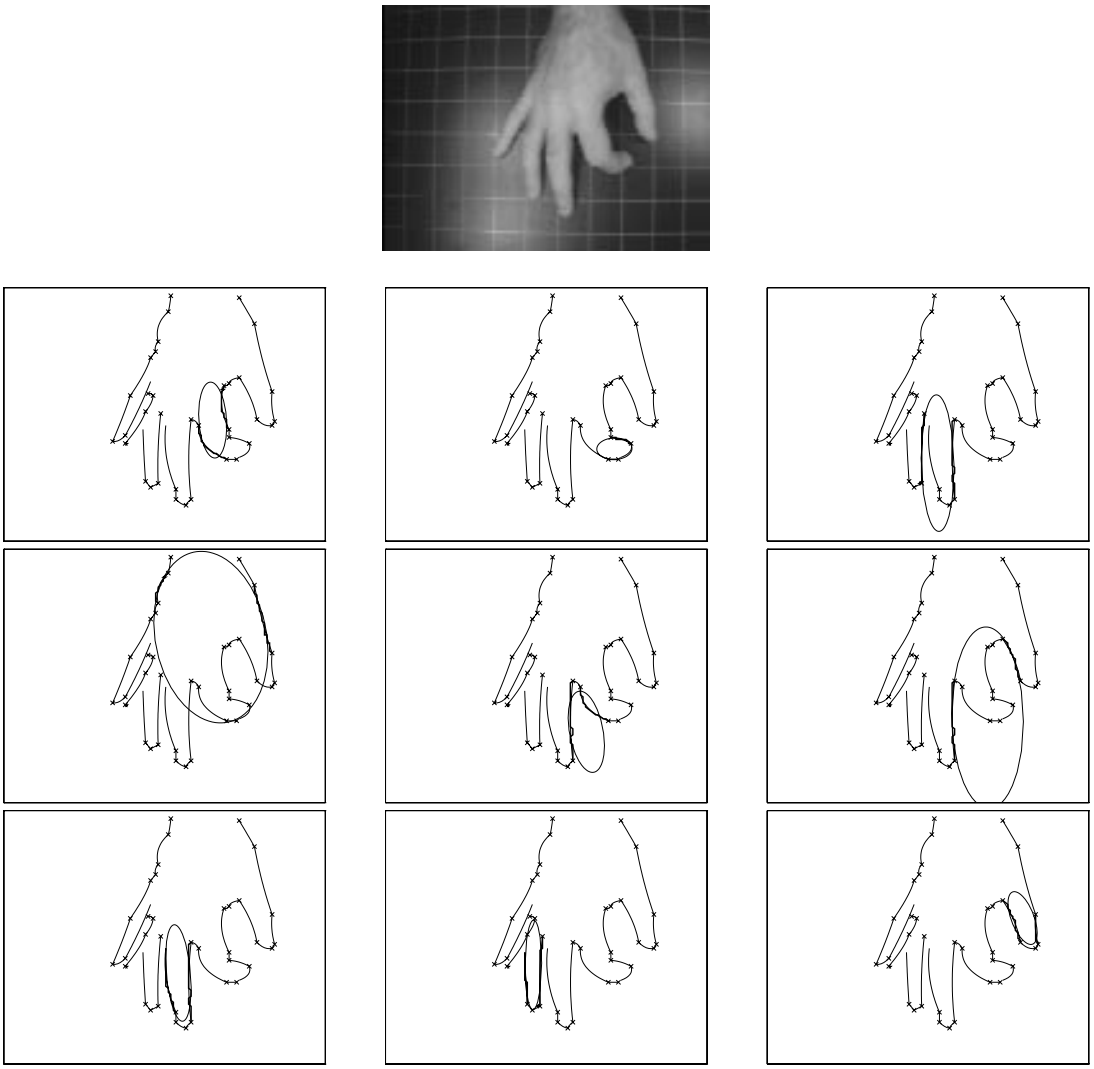


Figure 4.9: Nine initialisation examples to pairs of codons for hand test image. The pairs of codons are represented by the rugged lines. Good initialisations are produced for all the fingers. The back of the hand is not properly initialised. As an illustrative example, some initialisations to part-plausible (although not corresponding to actual parts) codon pairs have been included. Due to the problems described in Section 4.6.1, is some cases these initialisations can go quite wrong.

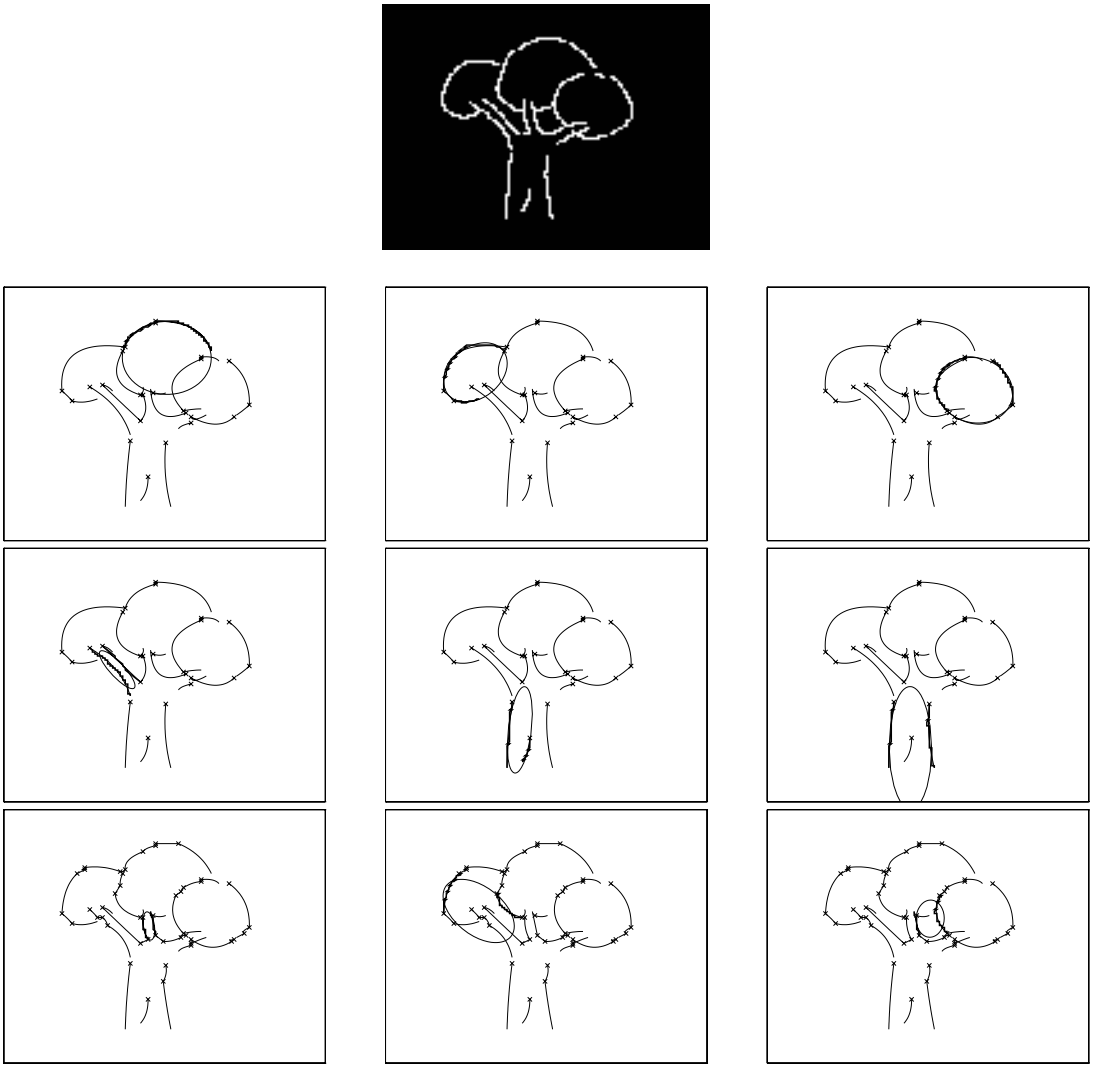


Figure 4.10: Nine initialisation examples to pairs of codons for tree test image. The pairs of codons are represented by the rugged lines. Good initialisation are produced for all the bushes but the trunk is slightly elongated.

4.6.2 Model pre-shaping

Once the coarse dimension, position and orientation of a part hypothesis are found through ellipse fitting to a seed group of codons as described in the previous subsection, the part-like PDM presented in Section 3.4 is *(i)* initialised by scaling, positioning and rotating accordingly, and *(ii)* fitted to the seed codons by the technique described in Section 3.4.5.

This phase is called *pre-shaping* and aims at easing the gathering of further image evidence, thereby allowing a proper final fitting to the whole set of supporting codons. For instance, if the actual part is slightly bent, in this phase the PDM would bend and thus more matching codons can be found along the PDM contour.

The fitting technique used in this phase is the same as the one described in Section 3.4.5 and therefore we shall not dwell on further details; the only important thing to say is that at this stage only a few iterations (about 10-15) are performed, since a precise fitting is not required as yet.

A few experiments are described here in pre-shaping to seed pairs of codons.

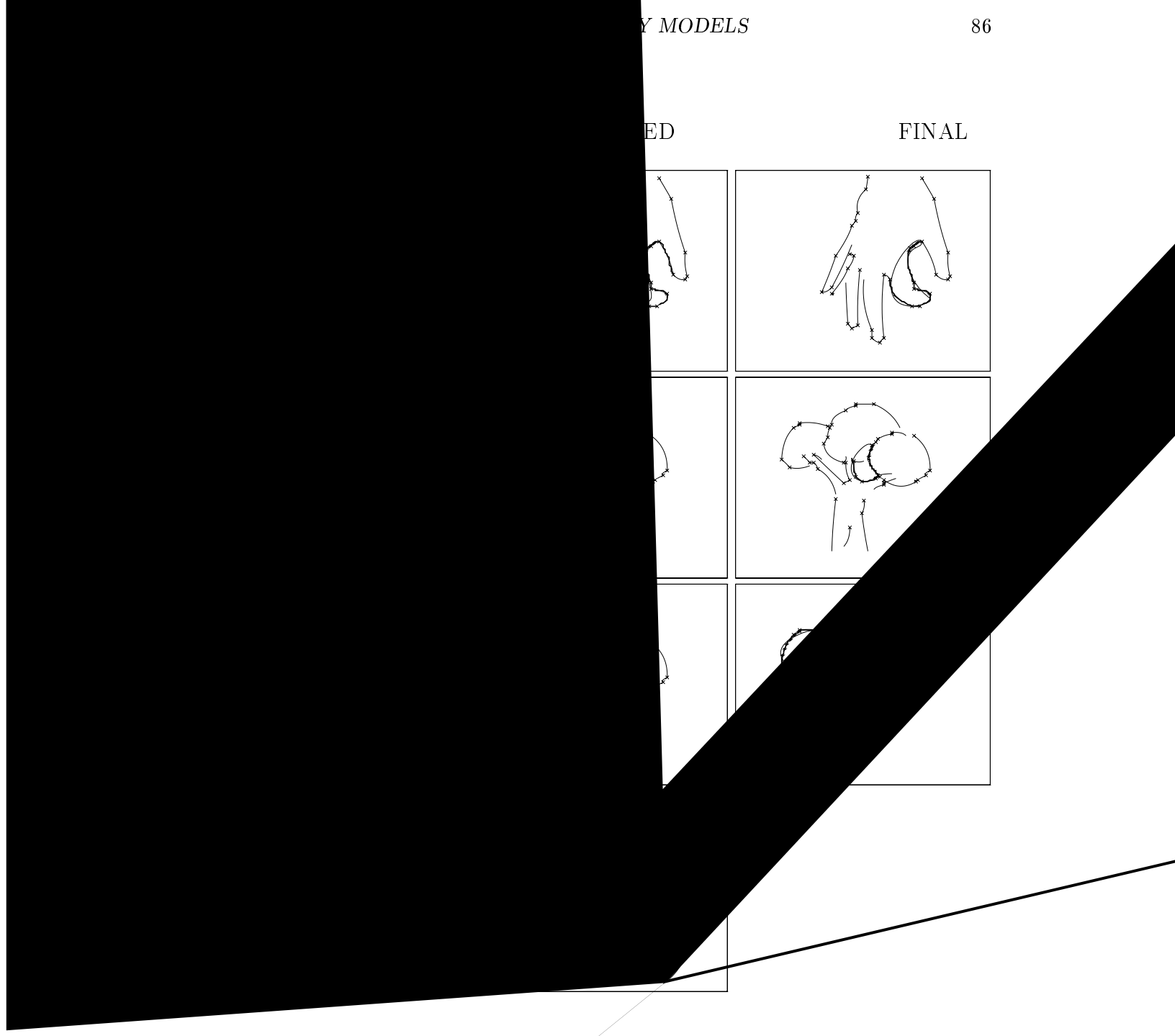
The middle column in Figure 4.11 shows five pre-shaping results to pairs of codons; the corresponding elliptical initialisations are shown beside each example. It might seem that in some cases the pre-shaping does not have relevant effects but it is indeed of crucial importance for initially registering size and position of the PDMs before the next global fitting stage can proceed.

4.6.3 Finding supporting codons

When a part hypothesis is available, e.g. after pre-shaping, it is necessary to ascertain its suitability by looking for further supporting codon evidence in the image.

In this subsection, the method used for determining supporting codons in presented; although it will be mainly used in the filtering stage given in the next chapter, it has been put here because the method is also loosely used for determining the neighbourhood of the pre-shaped models in order to perform the final fitting.

ED FINAL



Let us now introduce some formal notation.

Let $\mathcal{C}_j \in \mathcal{I}$ be a codon, that is a topologically connected set of n_j image pixels $p_1, \ldots p_m, \ldots p_{n_j}$ in the image $\mathcal{I}$. The aim is to determine a set of codons $\mathcal{R}_i = \{\mathcal{C}_{j_1}, \mathcal{C}_{j_2}, \ldots, \mathcal{C}_{j_k}\}$, called the *supporting region*, that are likely to correspond to (i.e. match) the hypothesis contour.

In works such as [Leonardis *et al.* 95] or [Darrell & Pentland 95], the support regions of model hypotheses are a *set of pixels* for which the Error of Fit (EoF) is less than a given threshold that depends on the noise level but this is not needed here, since pixels are already grouped into codons that belong to a single part, and therefore the support region is expressed in terms of which codons conform to the model hypotheses. In our context, the codon-hypothesis displacements are not Gaussian, and in fact not even of an *a priori* known distribution, which makes the choice of the EoF distance measure non-trivial.

The following *empirical* acceptance rule has been employed to decide whether a codon belongs to a model.

Let $d(\mathcal{H}_i, p_m)$ be the approximate *signed* geometric distance of a generic image point p_m to a PDM as described in Section 3.4.5 and let

$$\mu_1(\mathcal{H}_i, \mathcal{C}_j) = \frac{1}{n_j} \sum_{p_m \in \mathcal{C}_j} d(\mathcal{H}_i, p_m)$$
$$\mu_2(\mathcal{H}_i, \mathcal{C}_j) = \frac{1}{n_j} \sum_{p_m \in \mathcal{C}_j} (d(\mathcal{H}_i, p_m) - \mu_1(\mathcal{H}_i, \mathcal{C}_j))^2$$

be the first and second order moments of the displacements.

A codon $\mathcal{C}_j$ matches a model contour if:

$$\begin{cases} \frac{|\mu_1(\mathcal{H}_i, \mathcal{C}_j)|}{s} & \leq & \overline{\xi_1} \\ \frac{\mu_2(\mathcal{H}_i, \mathcal{C}_j)}{n_j} & \leq & \overline{\xi_2} \end{cases}$$

where $\overline{\xi_1}$ and $\overline{\xi_2}$ are two thresholds, s is the scale of the PDM (Section 3.4.4) and n_j is the length (in pixels) of the codon.

The first inequality aims at thresholding the average codon/model displacement and, since this quantity expresses an absolute value, it should depend on the scale of the model to give good discrimination. A good way of embodying this relationship is to

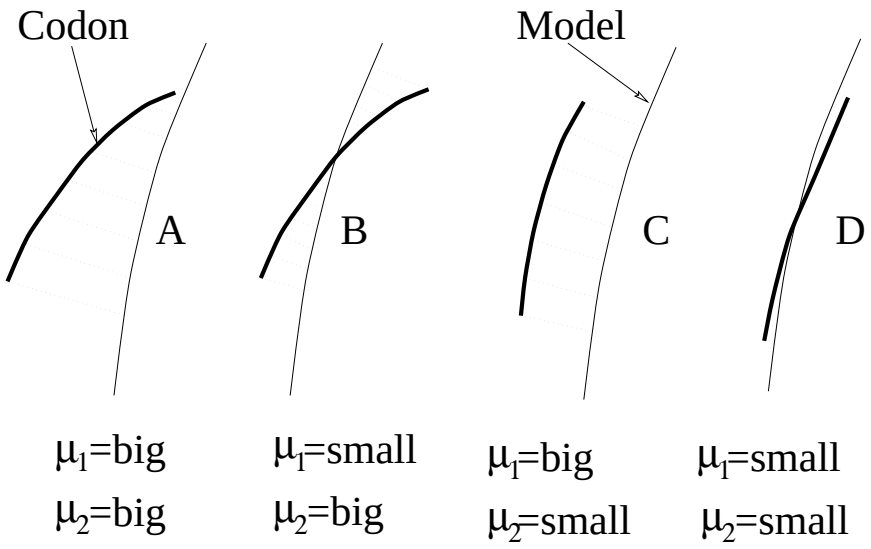


Figure 4.12: Qualitative taxonomy of codon-hypothesis displacements.

make the threshold vary linearly with the recovered scale s of the PDM (Section 3.4.4), therefore the division by s of the left-hand side of the first inequality.

Because the division by n_j of μ_2 has the effect of normalising the second moment with respect to the length, the second inequality thresholds a value proportional to the slant between codon and model contour.

Figure 4.12 illustrates typical cases of codon-model displacements. Case A shows a situation where the codon and model intersect with a certain degree of slant. In such a case, both μ_1 and μ_2 will be large. Case B shows a case where μ_1 is approximately zero but a high μ_2 helps identifying inconsistency; case C is the opposite of B, and a high μ_1 helps rejecting the codon. Case D happens when codon and hypothesised model are in accordance and both μ_1 and μ_2 are small. Cases A and C happen often in our framework and for this very reason the assumption of zero-mean noise is meaningless.

In [Pentland 90] and [Leonardis *et al.* 95] the pixel-wise EoF function was a function of the variance alone. This was possible because they assumed not only small deviation of the data from the model but also zero mean. However, generally speaking, a relatively high average displacement is allowed with respect to the slant angle: highly slanted codons are unlikely to be part of the models.

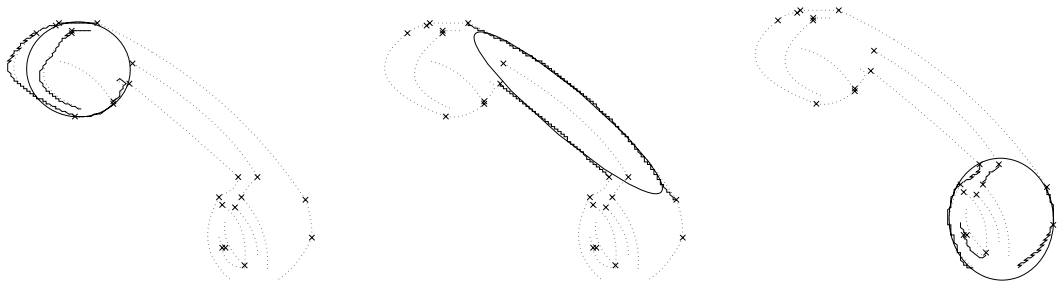


Figure 4.13: Example of supporting codons of three elliptical hypotheses (ragged lines).

The two thresholds can be chosen to allow coarse or fine selection of matching codons. In the model fitting of Section 4.6.4, since the PDM is just pre-shaped, it is necessary to have a large threshold in order not to exclude potentially matching codons. In the experiments, they are set to $\overline{\xi}_1 = 5$ and $\overline{\xi}_2 = 0.2$. Figure 4.13 shows three simple examples in which the ragged lines give the support regions found with these two values; notice the presence of overlapping supports, which are common with high thresholds. In the more refined phase of hypothesis filtering that will be discussed in the next chapter, models are well shaped; if they correspond to actual parts, small codon/model displacements are to be expected and, therefore, lower thresholds are used. In all the experiments that will be given in the next chapter, it is $\overline{\xi}_1 = 2$ and $\overline{\xi}_2 = 0.1$.

An important final remark is due. Because of our support finding method, the whole model fitting and filtering procedure (next chapter) can be seen as using a "censored" norm on the data, which is a typical action taken by robust estimation techniques, of which the model-guided grouping method presented in this thesis is a special case.

4.6.4 Final fitting

Once the model has been pre-shaped, further image evidence – i.e. supporting codons – is found as outlined in the previous subsection. Then, starting from the pre-shaped position, the PDM is iteratively fitted to all the supporting codons. The fitting procedure is again the same as given in Section 3.4.5 but a few remarks will be given later. Once the fitting has converged, the supporting codons are found that constitute a grouping hypothesis. As remarked previously, differently from other approaches

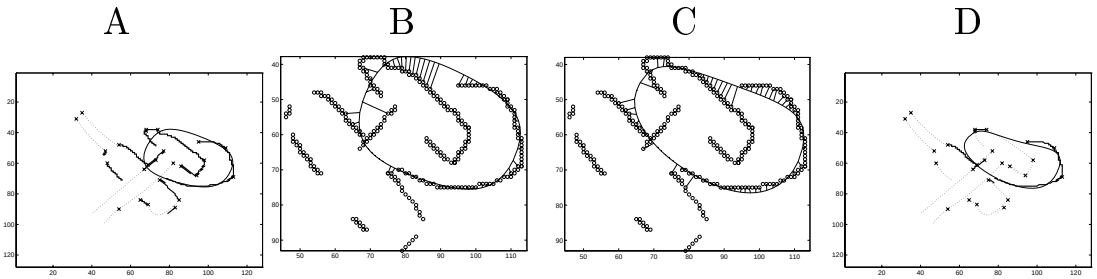


Figure 4.14: Model Fitting: (A): initial pre-shaped model with the selected neighbourhood; (B) and (C): two iterations in which the data points, the model and the point-to-point correspondence are shown; (D) the final result shown with the final supporting codons making up a part grouping hypothesis.

in our model-guided grouping strategy, grouped hypotheses are homologous to part hypotheses.

In the experiments that have been carried out, it has been noticed that a relatively low number of iterations (about 20) are necessary to obtain a good fitting to the data; for bent parts the fitting takes longer (roughly twice as long) but good convergence was always achieved except in cases where the codons did not provide enough support for the shape to be recovered.

A curious phenomenon that was noticed during the experiments was oscillations (or bouncing) of the model about the correct solution. To avoid such a problem, the weights w_t, w_θ, w_s and $\mathbf{w_b}$ of Section 3.4.4 are made to slowly decrease with the number of iterations, as also suggested in [Cootes *et al.* 94].

In order to allow the fitting to shapes that are considerably different from the initial PDM shape (that is a squarish ellipsoid, see the mid column of Figure 3.13), an extensive support region is found by setting rather large $\overline{\xi_1}$ and $\overline{\xi_2}$ in the support finding method as suggested in Section 4.6.3. As an example, Figure 4.14 shows the initial pre-shaped model with the selected large neighbourhood (A), two iterations in which the data points, the model and the point-to-point correspondence are shown (B and C), and the final result along with the final supporting codons making up a grouping hypothesis (D).

It can be seen that although the initial neighbourhood is quite large, global part-like

consistency is ensured by the use of the generic part PDMs. Although some points are attracted to other spurious features, their contribution is normally swallowed by the good ones, provided that enough part edges have been detected.

However, for highly cluttered images, the attraction to extraneous features may take over and convergence will not be achieved. I am currently investigating a new method for overcoming this problem – common to all model fitting schemes – by integrating in a single indivisible stage the powerful correspondence technique presented in [Scott & Longuet-Higgins 91] (see Section 3.4.5 and Appendix B) with a least squares PDM fitting. I have not yet found a solution to it but such a method might overcome most, if not all, the difficulties aforementioned.

4.7 Experimental results

In this section a number of examples of part-based grouping are discussed and results shown in Figure 4.15 to Figure 4.21. A larger set of hypotheses for each test image will be given in the next chapter in Sec. 5.2.

Where the raw edge images are given in earlier sections, they have not been included here. For some of the experiments, the initialisations were shown in Section 4.6.1. The codons are extracted, unless specified, with $d_{max} = 2$ for 128x128 images and $d_{max} = 4$ for 256x256 images. Along with the model, the ragged codons are the support found *after* the full model fitting by setting a low threshold as suggested in Section 4.6.3, that is $\overline{\xi}_1 = 2$ and $\overline{\xi}_2 = 0.1$.

In the first experiment shown in Figure 4.15, some good groupings of a synthetic 128x128 image of a beer bottle, a hammer and another roundish object are given. Of course, some groups do not correspond to actual parts but the most important thing is that those corresponding to actual parts are correctly recovered. It is worth noticing that, as it happens also in the other experiment in Figure 4.19, the grouping of the bottle body is recovered although it is occluded by the hammer handle: this feature is typical of our model-guided approach, which can overcome severe occlusions.

Figure 4.16 shows groupings of a real 128x128 image of a telephone handset. This image is considerably cluttered, with shadows, structural details and so on. There are

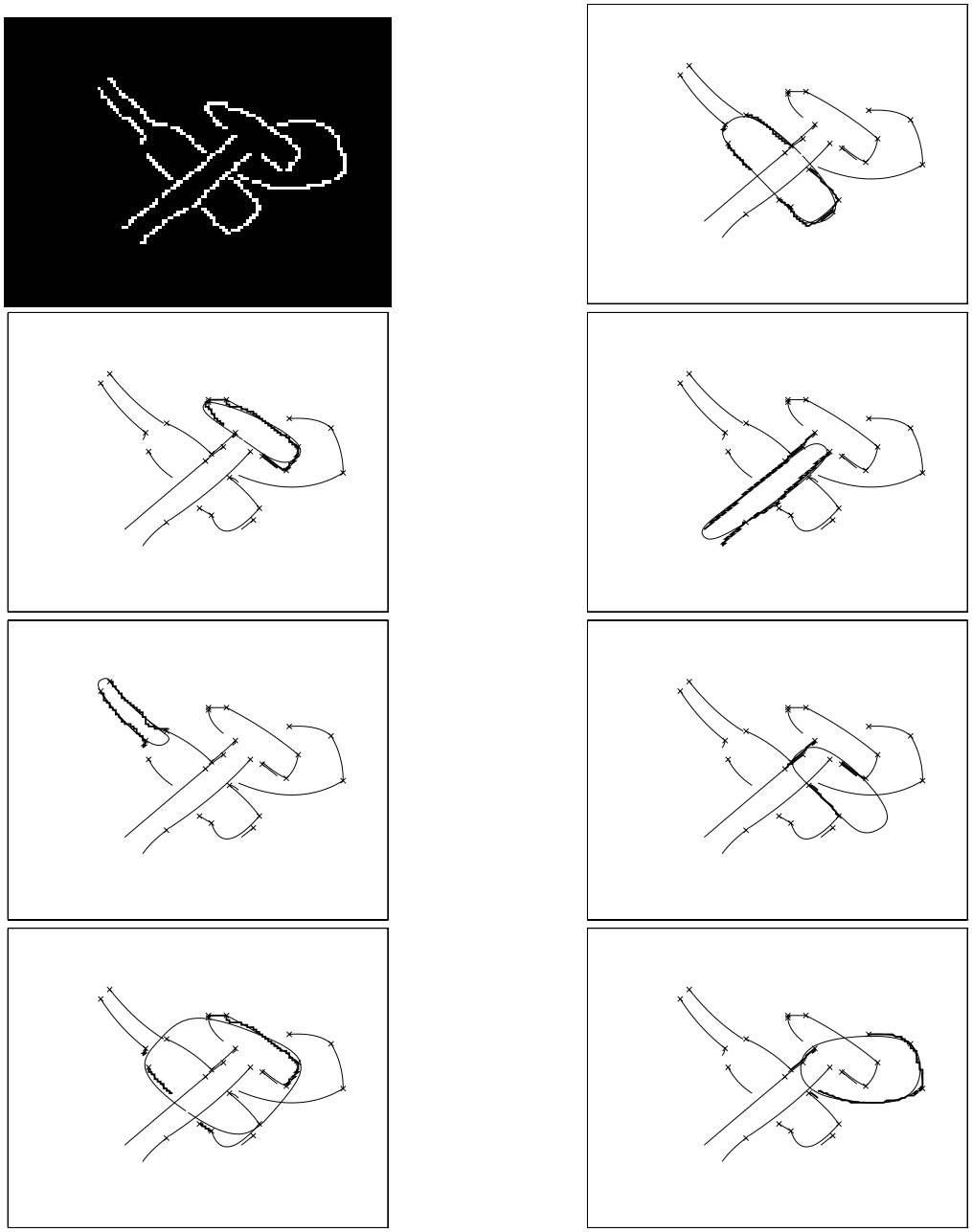


Figure 4.15: Some part groupings for the synthetic beer bottle and hammer example. Some groups do not correspond to actual parts but the ones corresponding to actual parts are correctly recovered. The grouping of the bottle body is recovered although it is occluded by the hammer handle.

HANDSET Example

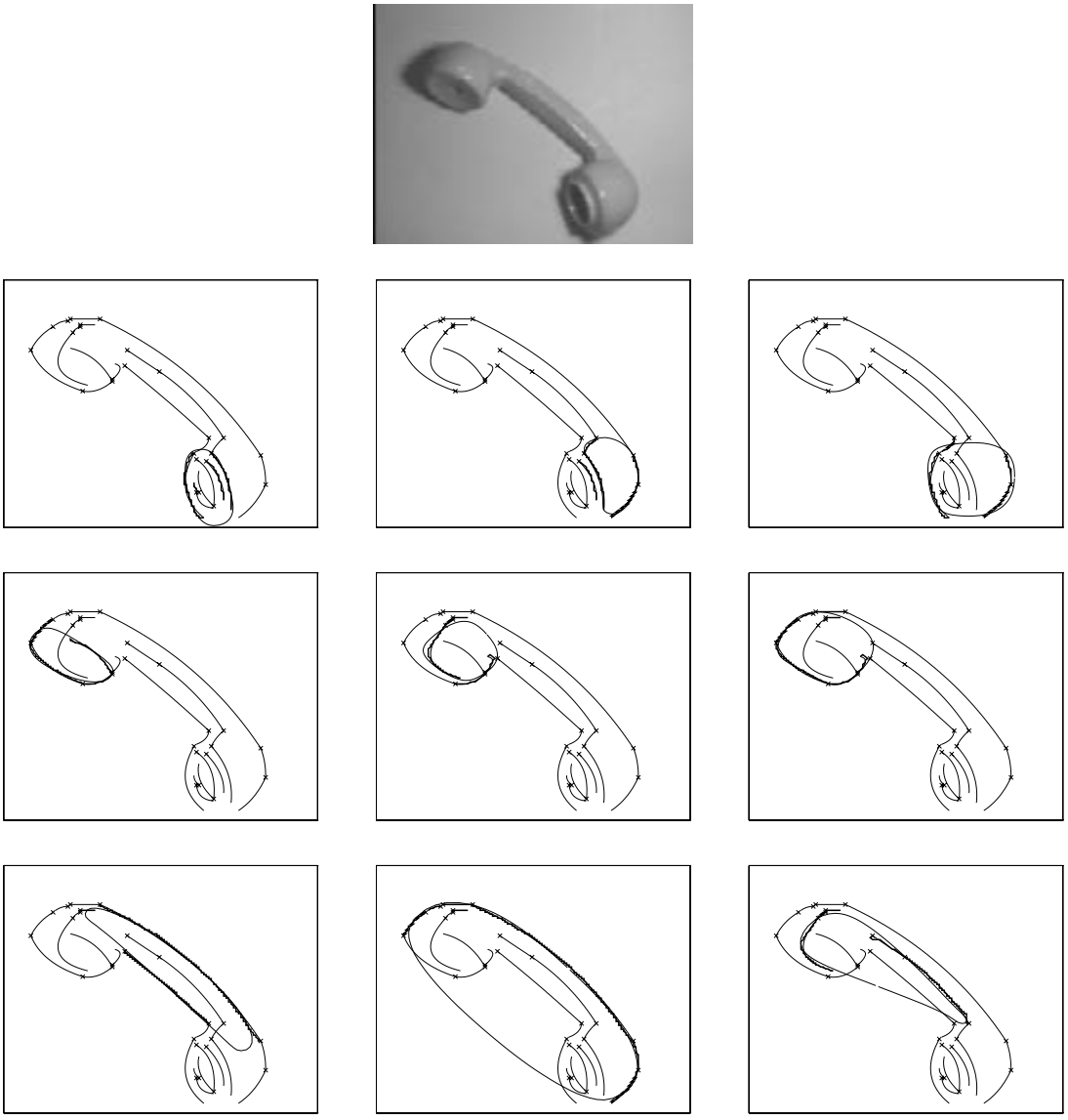


Figure 4.16: Some part groupings for the handset example. The original intensity image is in Fig. 3.8 and the edge image in Fig. 4.5. There are many good groups generated in this image, especially due to the circular rings in the ear (bottom) piece. Most of them will be filtered out, as shown in the next chapter.

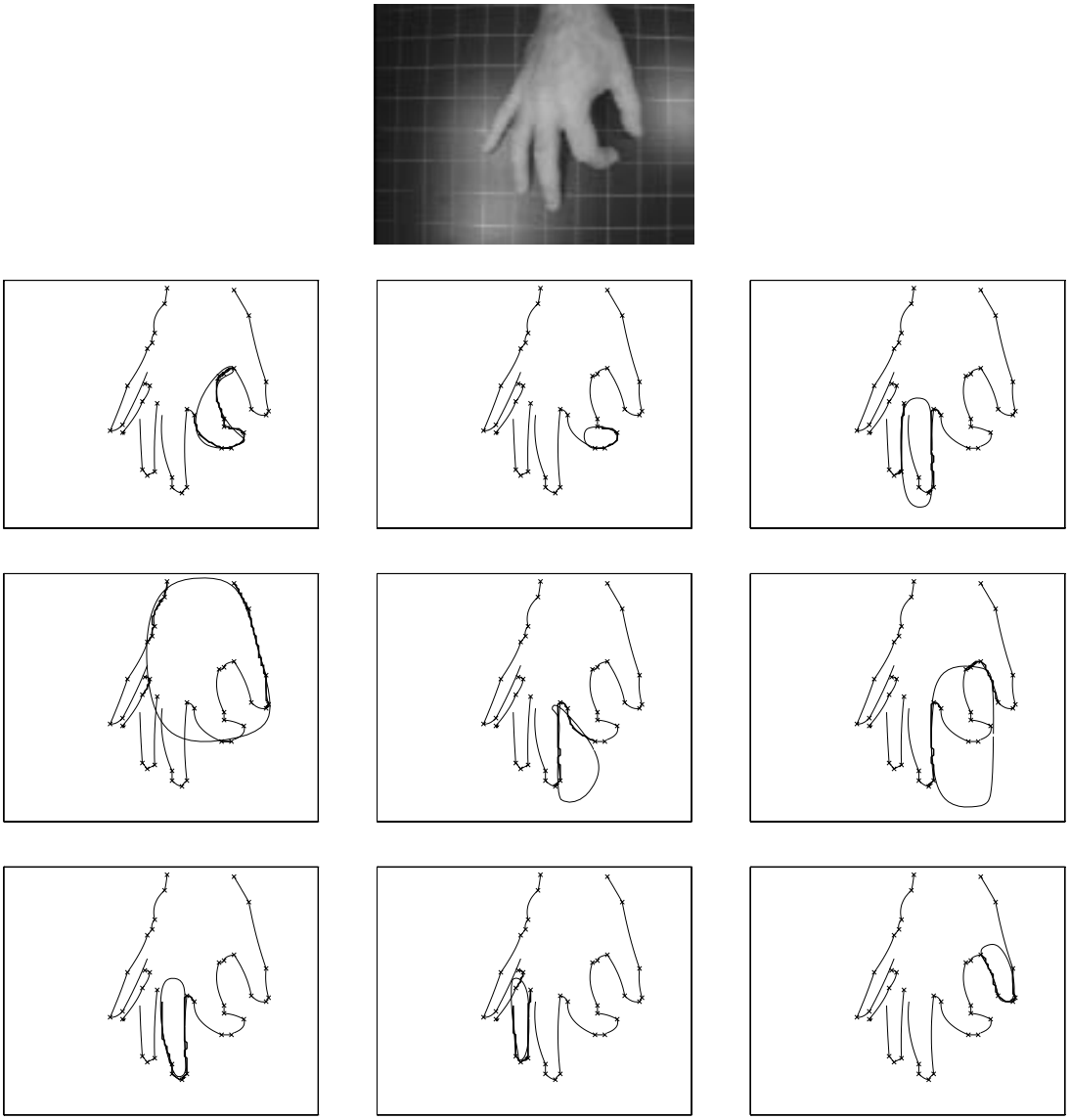


Figure 4.17: Some part groupings for the hand example. The original intensity image is in Fig. 3.8 and the edge image in Fig. 4.5. This is a pretty hard case, because the gaps between the fingers are all interpreted as possible part groupings: this is the classical figure-ground inversion problem. Moreover codons belonging to different fingers are often grouped together. The back of the hand has not been recovered for lack of codon support and bad initialisation.

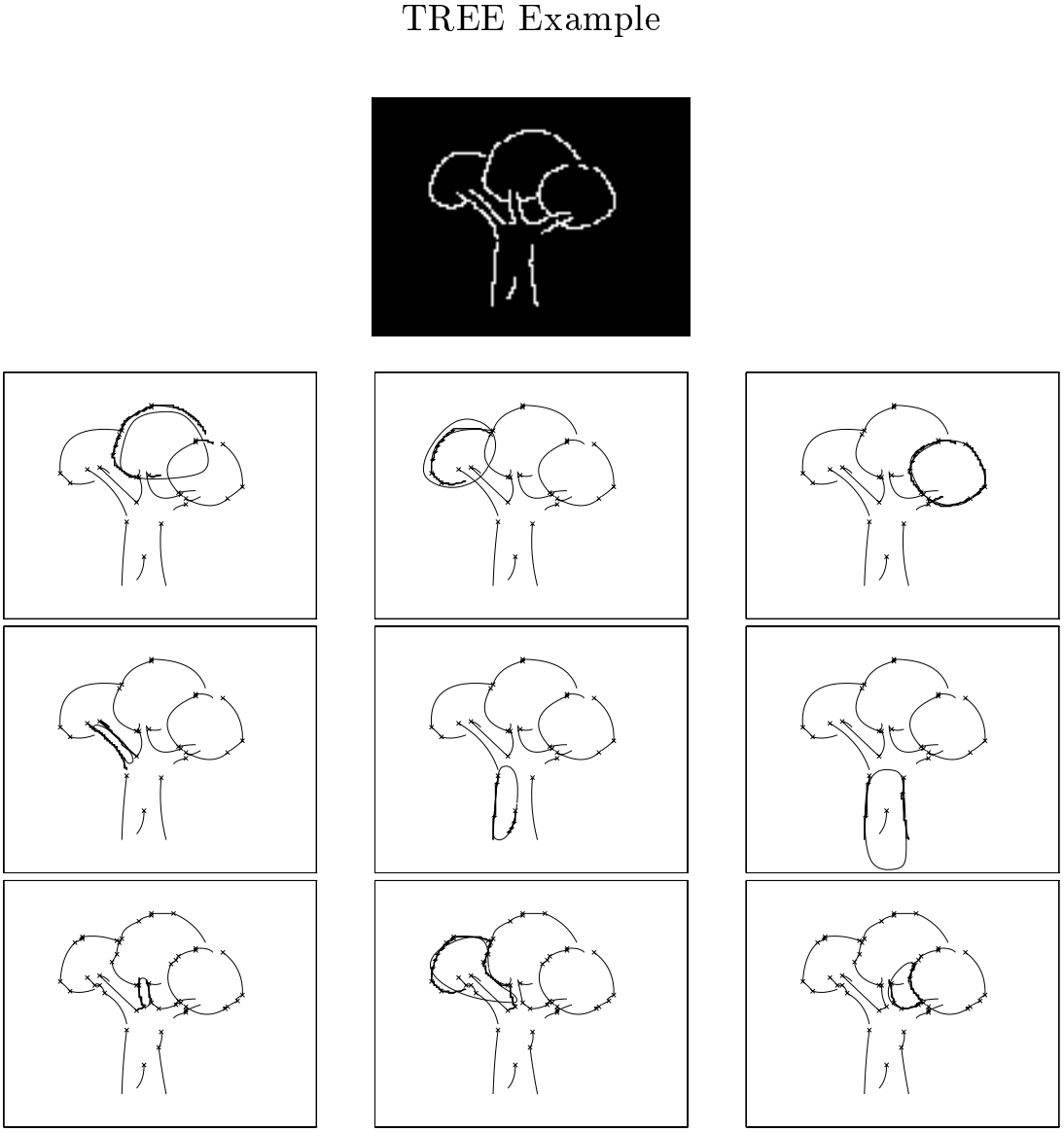


Figure 4.18: Some part groupings for the synthetic tree example. The three small branches are missed because the codon segmentation results too coarse for such small details; as an illustration, in the three bottom figures, the codon extraction scale was reduced to $d_{max} = 1$. See text for more details.

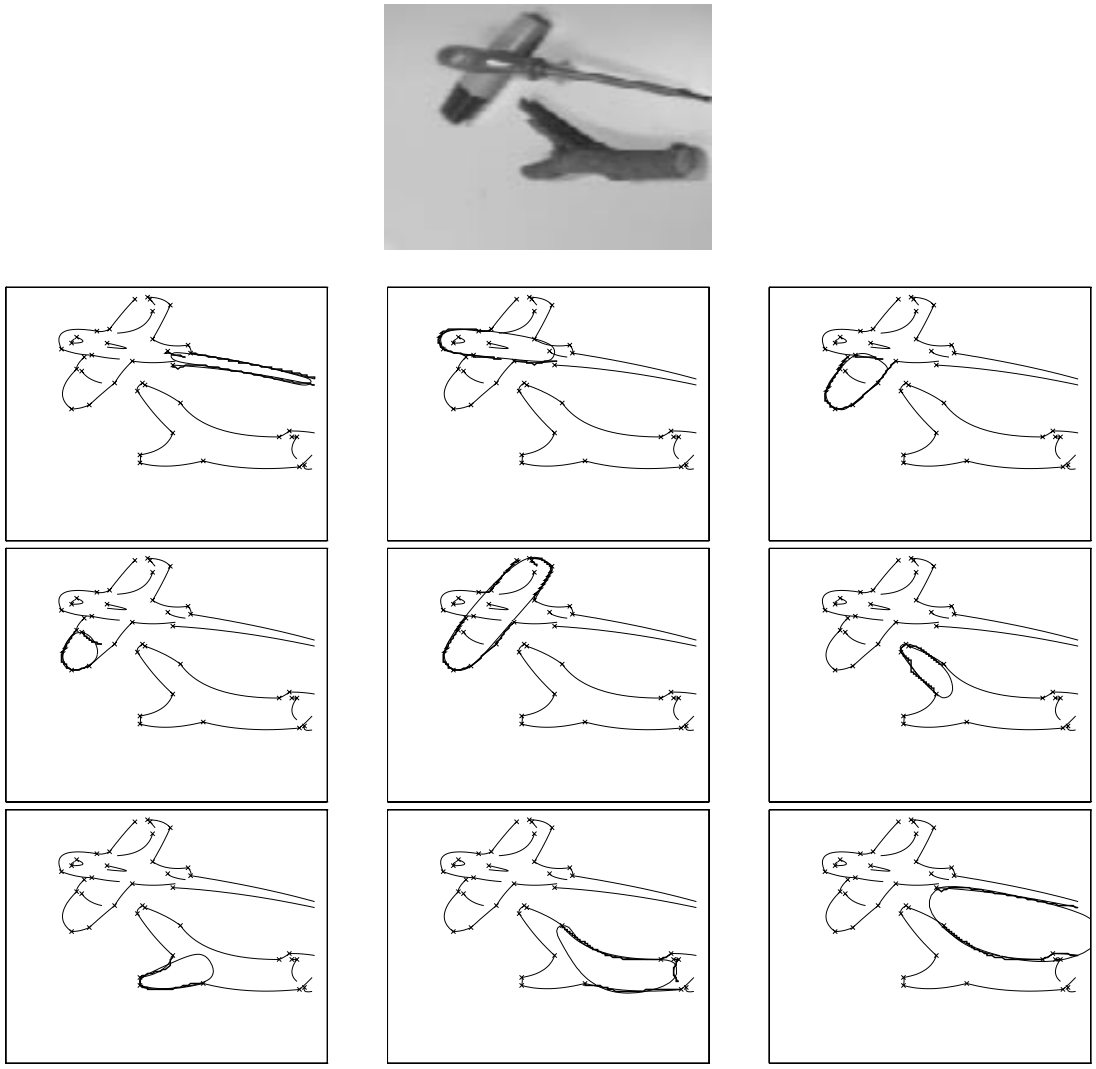


Figure 4.19: Some part groupings for the stick, marker and screw-driver example. The edge image is shown in Fig. 4.5. Parts are rather well defined here and, despite occlusion and cluttering, both the handle and the marker hypotheses are correctly produced.

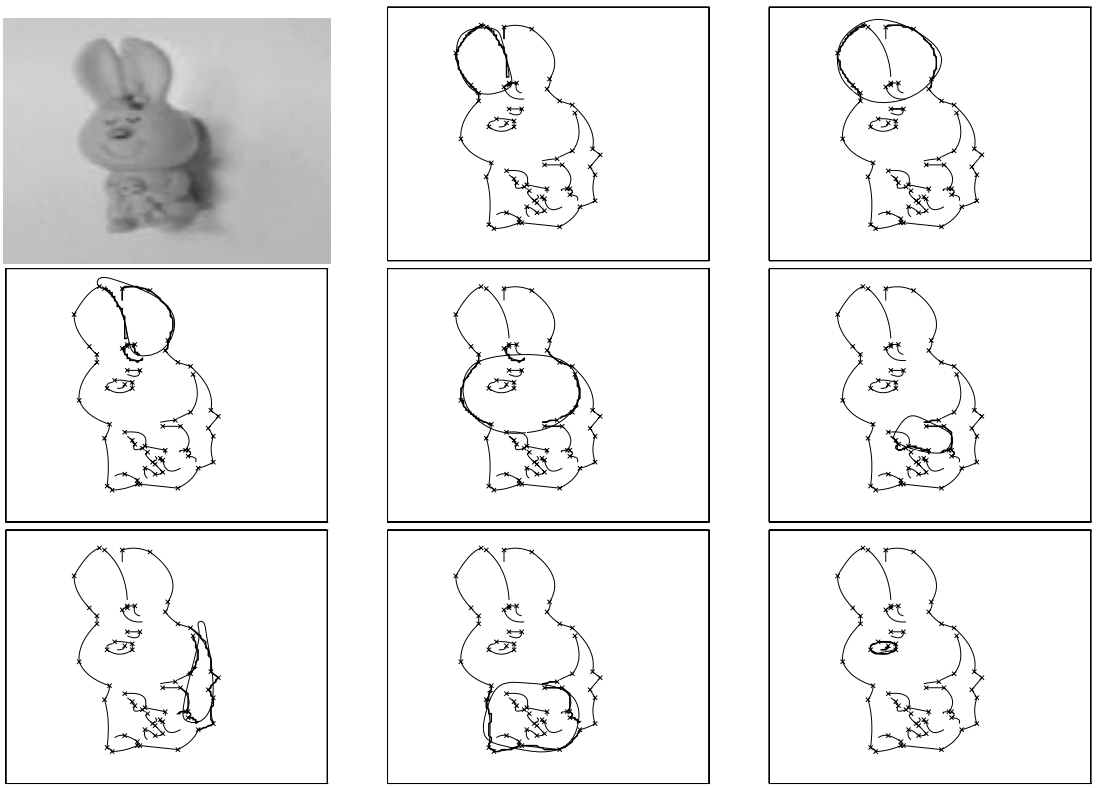


Figure 4.20: Some good part groupings for the toy rabbit example. All the correct main part groupings are found but, due to poor edge detection and resolution, the paws are not identifiable from the edge image.

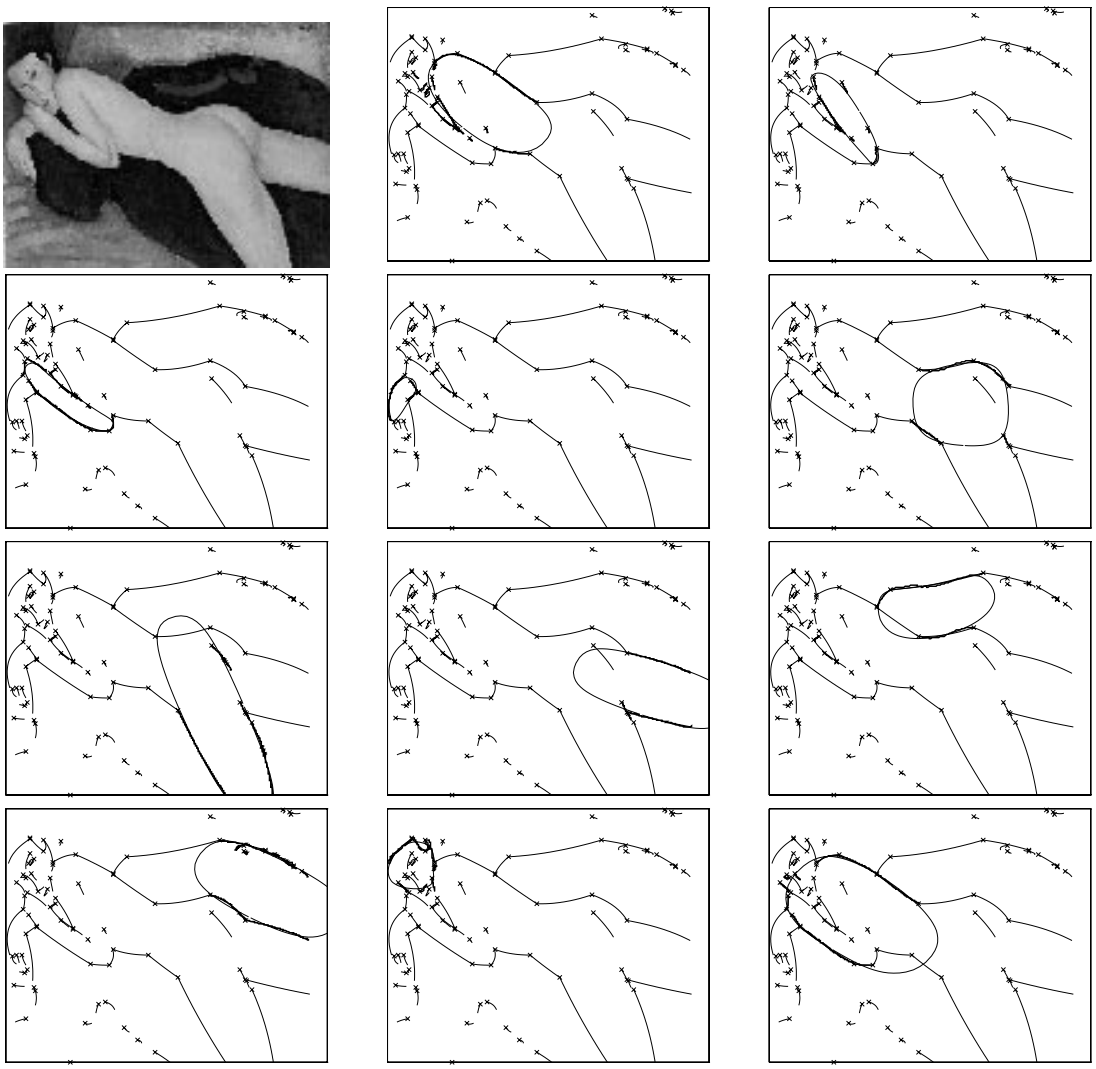


Figure 4.21: Some part groupings for the Modigliani's painting example. Apart form the rather elongated recovered model of the leg in front due to a bad initialisation, the correct groupings have all been recovered.

many good groups generated in this image, especially due to the circular rings in the ear (bottom) piece. Most of them will be filtered out, as shown in the next chapter.

Figure 4.17 shows some grouping examples with the 128x128 hand test image. All the correct finger groupings are recovered as shown in Fig. 5.4. However, this is a pretty hard case, because the gaps between the fingers are all interpreted as possible part groupings: this is the classical figure-ground inversion problem. Moreover codons belonging to different fingers are often grouped together. The back of the hand has not been recovered for lack of codon support and bad initialisation. As we shall see in the next chapter, the global filtering strategy that is employed will be able to easily discern the right grouping corresponding to the fingers.

A synthetic 128x128 tree example is given in Figure 4.18. This is an interesting example that shows how, if the appropriate scale is not chosen, some small parts can be missed out. The three small branches are missed because the codon segmentation results are too coarse for such small details; as an illustration, in the three bottom figures, the codon extraction scale was reduced to $d_{max} = 1$ and finer codon segmentation was achieved that allowed, for instance, the central branch to be recovered. As said in Section 4.4, this matter has not been taken into consideration in this work and a fixed d_{max} is used that depends only on the image resolution.

Figure 4.19 shows a 256x256 real image of a marker, a screw-driver and a wooden stick. Note how, as in the example of Figure 4.15, there is no problem in extracting the right grouping for the occluded marker.

The 256x256 toy-rabbit image in Figure 4.20 is another interesting example. All the correct main part groupings are found but, due to poor edge detection and resolution, the paws are not identifiable from the edge image.

Finally, Figure 4.20 show an experiment with a 256x256 image of a human figure painting, the "Nu couché de dos" by Modigliani. Apart form the rather elongated recovered model of the leg in front due to a bad initialisation, the correct groupings have all been recovered.

4.8 Discussion

In this chapter a new method for achieving part-based grouping through the use of simple part models has been presented. This has been achieved by first decomposing the image into codons and then by pre-grouping them into small sets that give enough structural information for the part model to be pre-shaped. Once the model is pre-shaped, a full-image fitting is performed that produces, along with a part-based codon groupings, part hypotheses.

Many hypotheses are produced by this stage and most of them are likely to be either meaningless or duplicates; this is due to the inherent ambiguities in edge images that can be solved only by a global analysis of all the hypotheses, which will be the subject of the next chapter.

In the following subsections, a discussion on the original contributions, the limitations and possible extensions is presented.

4.8.1 Contributions

This chapter contains several noteworthy contribution to the computer vision research:

- First above all, the part-based perceptual grouping is a new concept which is conceptually different from either convex grouping or the use of symmetry;
- All model-based part segmentation methods presented in the past rely on silhouette input, most notably the one in [Pentland 90]. Here this assumption has been dropped and a new computational method has been proposed that allows generation of part hypotheses from *real* edge images that are based on the intrinsic properties of codons; the method can be also extended to other domains, such as the segmentation into parts of range data images. The method presented in this chapter can also be used to significantly improve the the brute-force hypotheses generation method in the silhouette-based part segmentation work in [Pentland 90];
- The concept of model *pre-shaping* has been introduced for fitting deformable models to *unsegmented* image data. In the large majority of works the data are

assumed to be already segmented but this rarely happens in real world problems. The pre-shaping exploits the postulate that for simple deformable models, self-symmetries allow just a few image features to give sufficient structural information for the coarse shape to be recovered. This has the effect of focusing on a certain region of the parameter space before performing the full model fitting, thereby coping much better with extraneous features and missing data;

- The new robust ellipse fitting algorithm presented in Section 3.2 has shown its utility in the initialisation phase of the model fitting procedure; we reckon that it is likely to become a very popular general purpose ellipse fitting method.

4.8.2 Limitations

The part-based grouping strategy proposed in this chapter has some inherent limitations which do not, however, undermine the value of its contributions.

The biggest limitation (or criticism) regards the very use of models for perceptual organisation which should be a qualitative task by definition. However the generic part PDM that has been used here is just a *different* model than those used in other perceptual grouping works such as [Sha'ashua & Ullman 88] or the ribbons of [Mohan & Nevatia 92]. As claimed earlier in the chapter, generic part models are used with the purpose to drive the grouping to keep part-like consistency and cope with reasonable amount of occlusion and missing edges.

Another criticism pertains to the difficulties that can be encountered in the fitting stages, especially robustness to spurious codons and missing data. Although it has been shown through several examples that the proposed method is rather robust, clearly no claim can be made about its infallibility. For instance, in the hand example, the rather evident hypothesis corresponding the the back of the hand has not been generated. However, when possible misfits occur for very cluttered images, other model-free grouping methods would miss out good hypotheses just as well.

At a first glance, convex grouping methods such as [Jacobs 96] might appear superior and more general. That is certainly true, but the model-guided method has in principle lower complexity because it avoids the blind search of all possible combinations that

have global consistency and can deal with non-convex parts.

Groupings of odd-shaped or very bent parts, like the handle of a tea-pot [Zerroug & Nevatia 94], are difficult, if not impossible, to recover due to representational and fitting limitations of the models under this grouping scheme. Symmetry-based grouping, such as [Mohan & Nevatia 92] can easily deal with these situations, as long as a clear edge image is available. However one of the aims of the work was to explore more global methods, which are able to more naturally deal with occlusion and missing contour portions.

4.8.3 Possible extensions

First above all, the codon extraction phase could be improved by taking into account regions of high curvature and by introducing an appropriate scale selection method. This would both reduce the number of generated hypotheses and produce better groupings.

As suggested in Section 4.6.4, the model fitting could be greatly improved by integrating correspondence and fitting by singular value decomposition; this would also overcome small problems sometime encountered in the initialisation of the PDMs, currently performed by ellipse fitting. These matters are being investigated.

Actually, the geometrical models could be dropped in favour of a more flexible and efficient method for finding support that would use a part-oriented perceptual organisation criteria. It is not yet clear how this can be done fast and reliably in cases other than the one with convex objects, as in [Jacobs 96]. In this regard, I have started investigating a statistical method that would use the same kind of training as the one employed for generating the PDMs as in Section 3.4.2

Finally, the fitting to image data might be more efficiently produced by taking into account other information, such as brightness, as is normally done for PDMs (Section 3.4.4). The pre-grouping phase would still produce initial hypotheses and then, instead of performing the final fitting to additional supporting codons, it could be performed as usually done to raw images [Cootes & Taylor 92]. The exploration of this avenue is left for future work.

Chapter 5

Part Hypotheses Filtering

5.1 Introduction

The previous chapter described how a redundant set of part grouping hypotheses is generated from a 2D edge image. In this chapter we discuss issues concerning the filtering of a set of hypotheses to retain only those that are likely to correspond to actual parts. Since the implementation of a sophisticated filtering method that accounts for more complex structural properties of the edge image would be a big research topic in its own right – therefore beyond the scope of this work – two *low-level* methods are presented here; namely filtering by perceptual salience thresholding and by support competition.

The first method (given in Section 5.2) sorts hypotheses according to a simple measure of perceptual saliency that mainly accounts for the percentage of PDM contours supported by edge data. Although the highest scoring hypotheses often – but not always – correspond to actual object parts, the set of hypotheses thus produced is frequently still redundant.

To overcome this problem, a significant method to produce a minimal set of part hypotheses is presented in Section 5.3. This is an extension of a recently developed successful segmentation technique based on the Minimum Description Length criterion which are used in [Leonardis *et al.* 95] and [Darrell & Pentland 95] to segment range data into 3D patches; here, its basic principles are for the first time applied to the segmentation of geometric primitives from real 2D images. Although good results have

been obtained, some principled limitations have been pointed out that had not been mentioned in previous works and a full account of them is given in Section 5.3.7.

The integration of other information, in particular background knowledge, to solve ambiguous cases is briefly discussed in Section 5.4 and an experiment is presented.

Finally, the chapter is concluded with a discussion on the contributions and the proposal of future work.

5.2 Sorting hypotheses by perceptual saliency

In this section, a simple saliency measure is defined that allows one to sort and filter part grouping hypotheses.

The *contour covering ratio* is a simple, yet significant, salience measure that has been used in some previous works. In [Jacobs 96], the only convex groupings that are considered salient, have over a certain percentage (70% in the paper) of their polygonal convex hull supported by edge data. In [Mohan & Nevatia 92], a saliency function that combines several measures includes the percentage of support of the two ribbon's sides.

As far as part segmentation goes, other salience criteria are possible, especially the use of local structural features, such as "T" junctions as in [Bergevin & Levine 93] or [Zerroug & Nevatia 94]. This kind of information is of rather high perceptual relevance but, due to objective difficulties in recovering it from real imagery, thus far its use has been restricted to artificially controlled images.

As repeatedly stated, one of the main aims of this work is to operate with real edge images and therefore techniques based on local non-accidental properties *alone* were not considered viable for our purposes because of their unreliability. However, it is clear that the integration of such information in a *coherent* manner with other perceptual clues will constitute a significant step ahead (see, e.g., [Sakar & Boyer 93]).

5.2.1 Definition of a perceptual salience measure

As shown in the previous chapter, the strategy of using part-like models to group codons results in a relatively small set of part hypotheses. Although most of them do

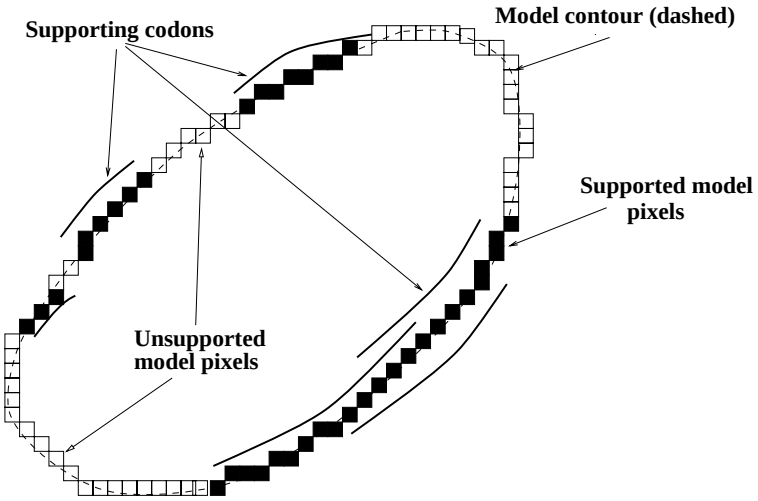


Figure 5.1: The supported model pixels are those subtended by the model supporting codons; they can be seen as model landmarks having a correspondence in the image data. Unsupported model pixels do not have correspondence in the image evidence.

not correspond to actual parts, the correct ones must have a certain amount of their contour supported by codons and thus, following [Jacobs 96], the contour covering percentage is considered here as the main perceptual clue.

Let us indicate by $\mathcal{R}_i$ the set of supporting codons of $\mathcal{H}_i$ determined by the method in Section 4.6.3, by $|\mathcal{H}_i|$ the number of pixels subtended by the model contour and by $|\mathcal{R}_i|$ the number of pixels subtended by the set of codons $\mathcal{R}_i$. Moreover, let us denote by $|\mathcal{R}_i \dashv \mathcal{H}_i|$[1] the number of pixels on the model contour that are *supported* by the codon evidence $\mathcal{R}_i$, which is in general different from $|\mathcal{R}_i|$.

In the pictorial example of Figure 5.1, $|\mathcal{R}_i \dashv \mathcal{H}_i|$ are represented by the black pixels whereas the sum of white and black pixels is equal to $|\mathcal{H}_i|$.

The covering ratio $S_C(\mathcal{H}_i, \mathcal{C})$ of a model hypotheses $\mathcal{H}_i$ given the set of codons $\mathcal{C}$ is defined as:

$$S_C(\mathcal{H}_i, \mathcal{C}) = \frac{|\mathcal{R}_i \dashv \mathcal{H}_i|}{|\mathcal{H}_i|}, \tag{5.1}$$

A wise choice of the error thresholds (Section 4.6.3) for the determination of the support

[1] The *symbolic* notation $y \dashv x$ indicates the elements of x that are related or corresponding y. In model matching, for instance, x could be a model and y image features and $y \dashv x$ indicates the set of model features that have correspondence in the image. Mnemonically, it can be interpreted as "projection".

ensures that likely supporting codons are selected and bad ones are always rejected. This choice is a very hard step because the model fitting is never precise and therefore codons that match the model poorly might nonetheless belong to the actual part. To overcome this problems there is no solution other than having a rather large threshold, as pointed out in Section 4.6.3.

In [Mohan & Nevatia 92] other heuristic measures, notably skew angle and aspect ratio, were used to increase the discrimination power of a perceptual saliency measure of ribbons that is solely based on covering percentage. Although our "non-rectangular" domain makes the use of skew rather inappropriate, the aspect ratio has a certain amount of relevance. The aspect ratio $S_{AR}(\mathcal{H}_i)$ is defined as the ratio between the major and minor axis of the PDM, which is a function of the modes b_1 and b_3 and the model scale as given in Section 3.4.

The salience measure $S(\mathcal{H}_i, \mathcal{C})$ adopted here is then a weighted sum of the covering percentage and aspect ratio:

$$S(\mathcal{H}_i) = S_C(\mathcal{H}_i, \mathcal{C}) + W_{AR} \cdot S_{AR}(\mathcal{H}_i) \tag{5.2}$$

The weight constant W_{AR} is rather small (0.05 in the experiments) so that the aspect ratio does not dominate the saliency measure, but a slight bias to more elongated primitives is produced.

5.2.2 Experiments with saliency thresholding

In this section some experiments with filtering by perceptual saliency thresholding are presented. Most of the comments on each experiment are in the figure captions, so here just an overview is given.

For each experiment, the whole set of hypotheses is displayed first, followed by the ones whose saliency is greater than a certain threshold, which is indicated at the top of each figure. The thickness of the contours indicates the relative salience of each hypothesis, that is the thinner ones have a saliency close to the threshold.

Since, due to possible quasi-identical initialisations in the pre-grouping phase (as described in Section 4.5), there are some duplicate part hypotheses but only the best

of these is retained. The similarity criterion is simply checking as to whether two hypotheses share more than 80% of their support.

In the first five experiments (the tree, screw-driver, handset, beer bottle and handset images) the same threshold of 0.6 allowed all actual part hypotheses to be selected along with a small number of spurious ones that have good image support. In cases such as the tree (Figs. 5.2-5.3), the hand (Figs. 5.4-5.5) and screw-driver (Fig. 5.6-5.7) some high-salience hypotheses are actually caused by figure-ground ambiguity. Occluded hypotheses have also obtained good salience, as shown in Fig.5.7 and 5.11, but if the occlusion were more pronounced this would not have happened; contour completion techniques such as [Williams & Jacobs 95] could be employed to overcome these problems. In the case of the handset (Figs. 5.8-5.9), there are several salient groups due to the pronounced tridimensionality of the image, a large shadow edge at the top piece and many structural details at the bottom piece - as shown in the edge image of Fig. 4.16. All the actual parts do well but many hypotheses not corresponding to any physical part score the highest.

In the more complicated cases of the toy rabbit and the Modigliani painting example of Figures 5.12-5.13-5.14 and 5.15-5.16-5.17, respectively, two results with different thresholds are shown. Undoubtedly, the results here are less exciting because there are too many edges arising from small details (in the rabbit body and in the face of the painting) and because the hypotheses corresponding to the two legs of the painting subject have been misfitted with large models (due to limitations in the initialisation stage as pointed out in Section 3.4). This latter problem could be easily solved by employing (or integrating) symmetry information.

Clearly, no information about global coherence of the resulting set of hypotheses is embodied in the method, since their goodness is a purely local property. Moreover, a particular choice of the salience threshold could cause either many good hypotheses to be missed or too many bad ones to appear.

In the next section, a novel global approach inspired by the Minimum Description Length criterion is presented that will explore a possible way of overcoming these limitations and producing a minimal solution.

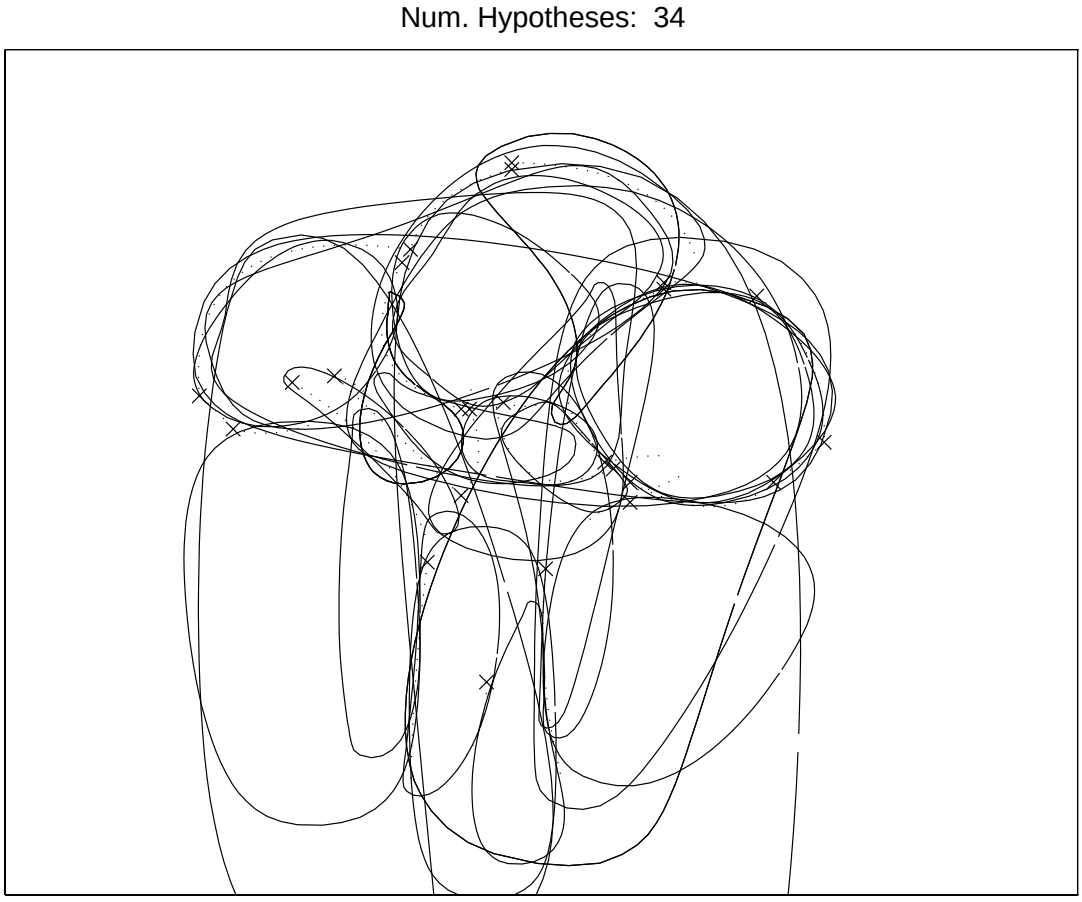


Figure 5.2: Set of part hypotheses for the tree example. The edge image and the codons can be found in Figure 4.18.

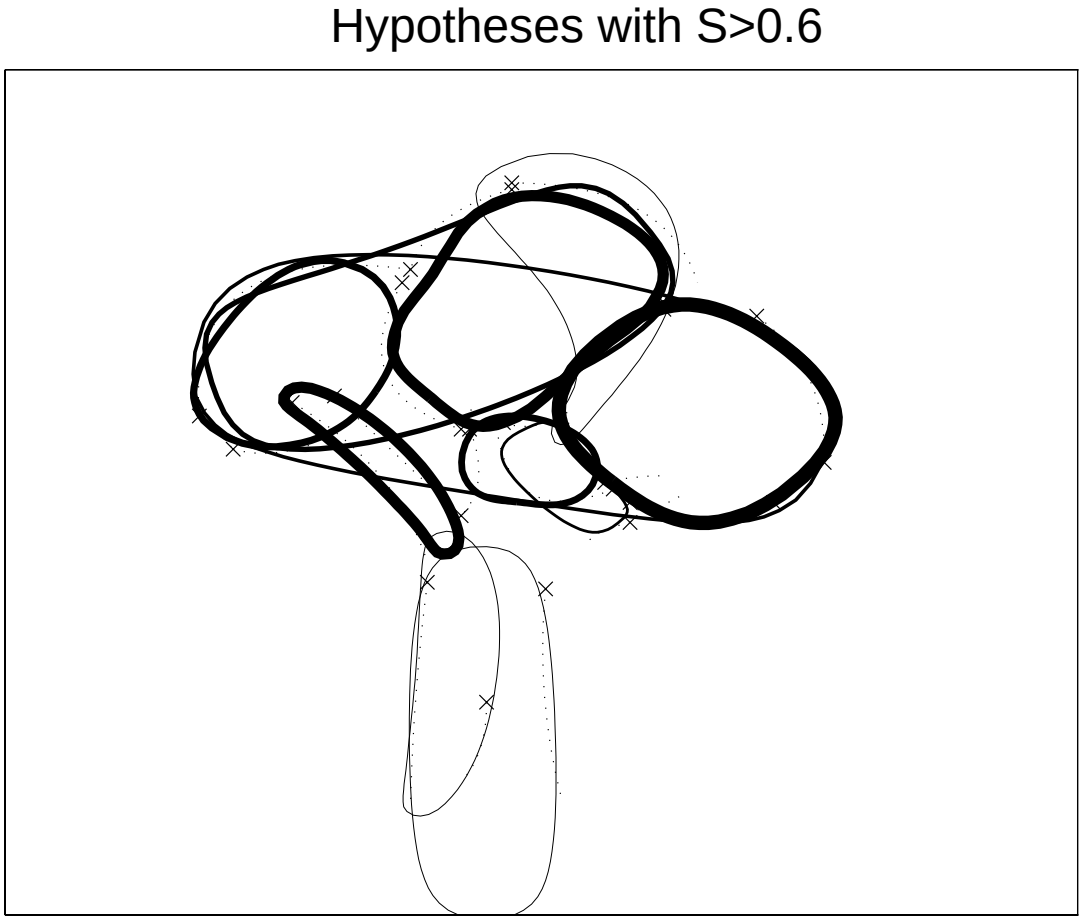


Figure 5.3: Hypotheses filtered by perceptual salience for the tree example. The central and right branch are not recovered for reasons of scale. The bushes have quite high salience. The trunk has low salience because the model fitting has produced too big a model due to lack of ends but it could be enforced by exploiting the high symmetry of the two delimiting codons. Note the two slightly elongated hypotheses encompassing distinct bushes.

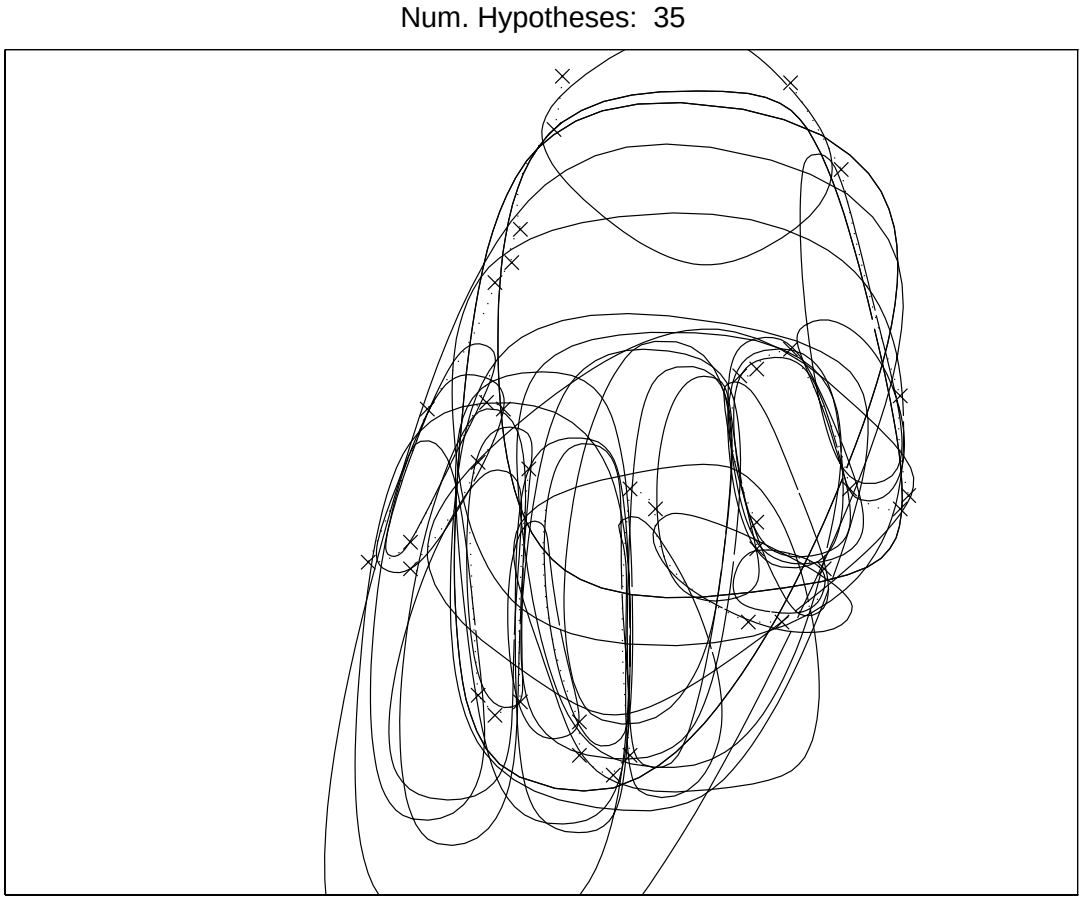


Figure 5.4: Set of part hypotheses for the hand example. The edge image and the codons can be found in Figure 4.17.

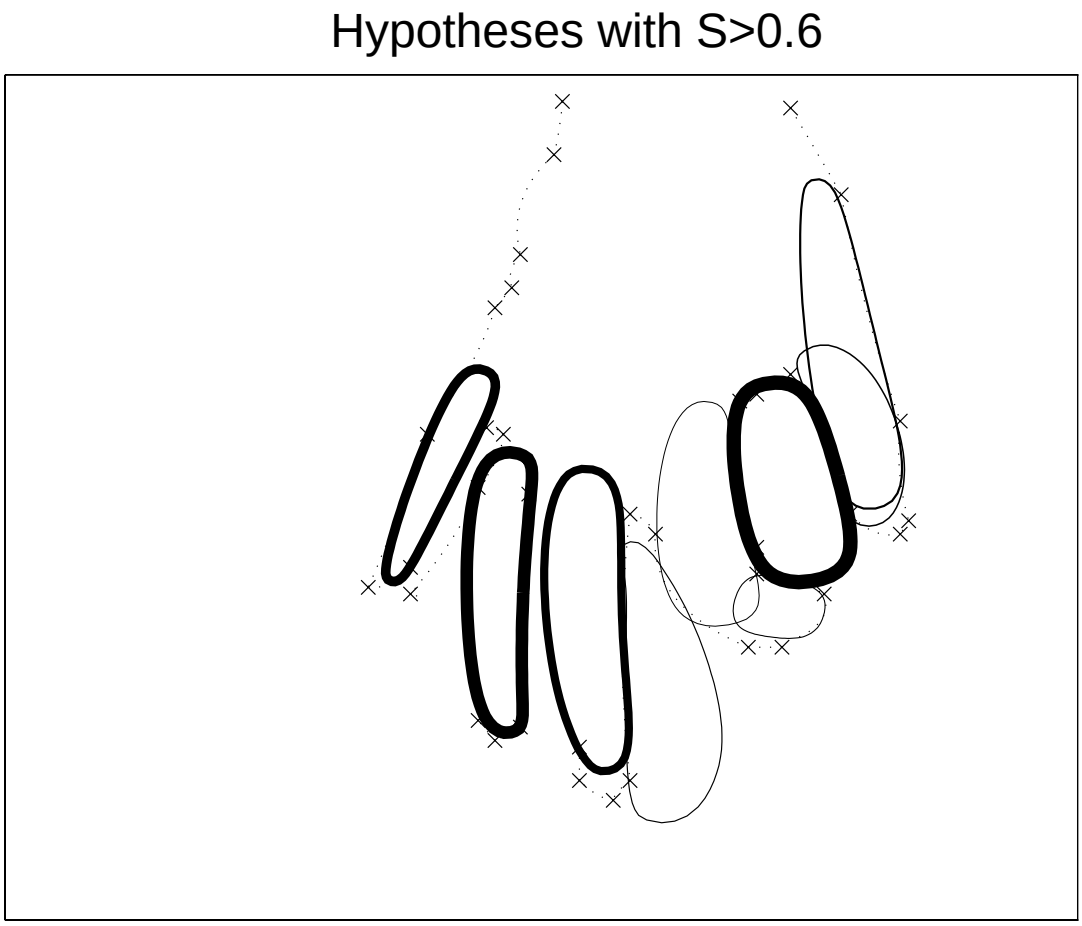


Figure 5.5: Hypotheses filtered by perceptual salience for the hand example. The little, ring and middle fingers have quite high scores ($S > 0.85$). The index and thumb, however, have lower salience due to lack of codon evidence. Note the very high score obtained by the gap between thumb and index caused by remarkable figure-ground ambiguity. The back of the hand, although well represented in the set of hypotheses shown in the previous page, does not have enough contour to have a high salience, the value is about 0.4.

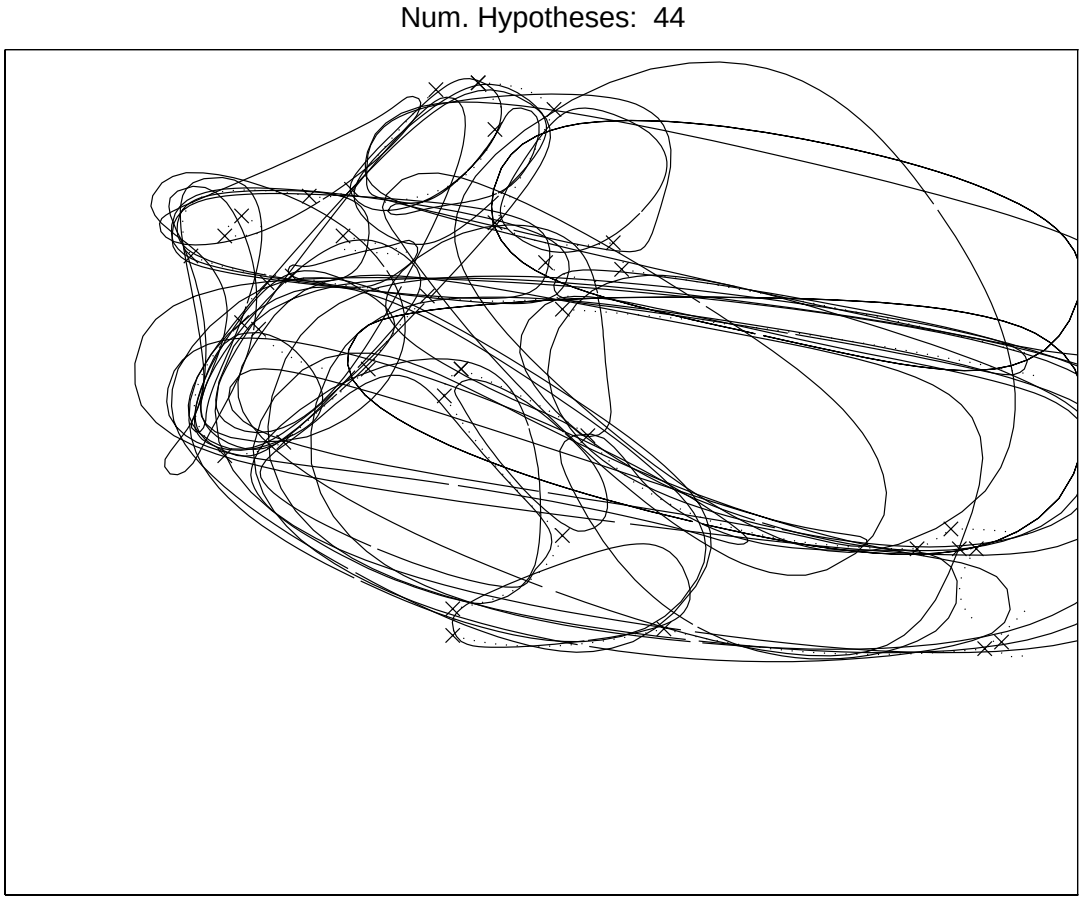


Figure 5.6: Set of part hypotheses for the screw-driver, marker and stick example. The edge image and the codons can be found in Figure 4.17.

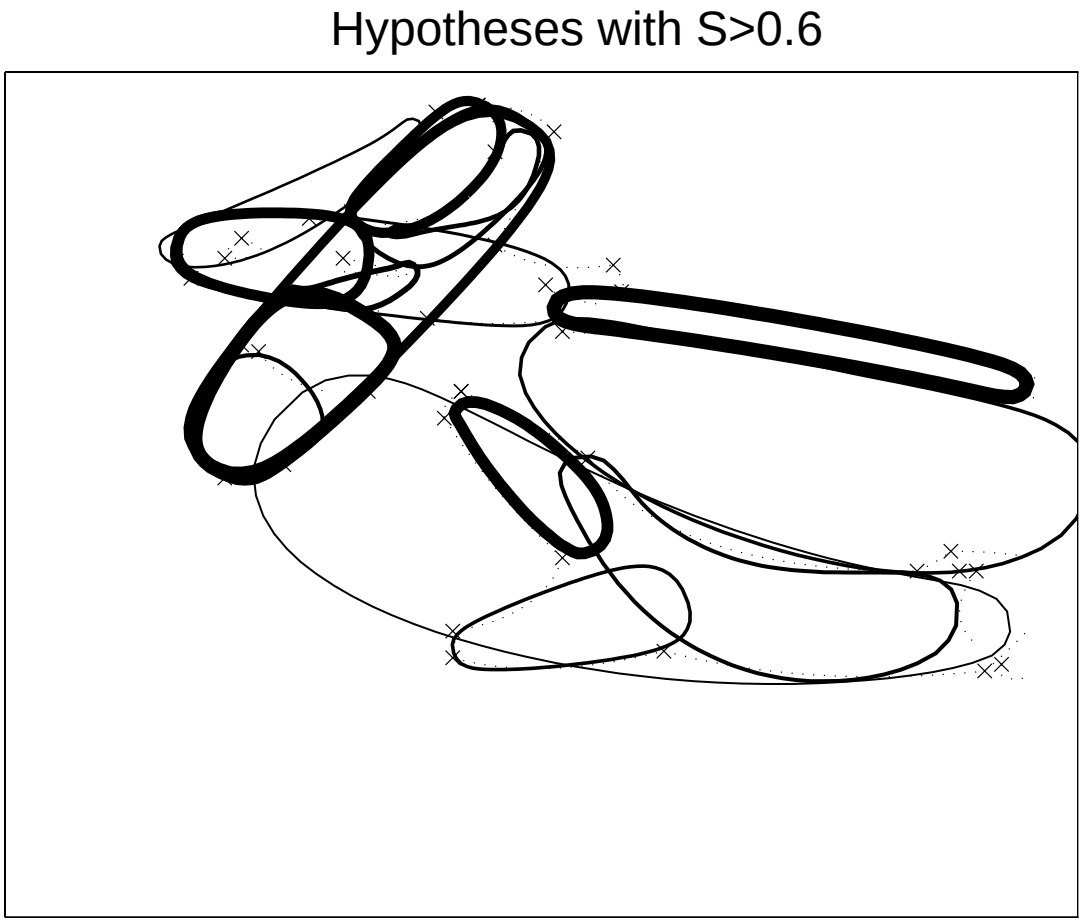


Figure 5.7: Hypotheses filtered by perceptual salience for the screw-driver, stick and marker example. The highest scoring hypotheses are the shaft, the top end of the wooden stick, the whole marker and some spurious ones originated by highly salient marking or occluding edges (see Fig.4.17). All the actual parts have good scores. Notice the big elongated shape that encompasses the whole wooden stick and the one bridging the top side of the stick and the shaft of the screw-driver: these have high salience too and only the use of more information could help disambiguate.

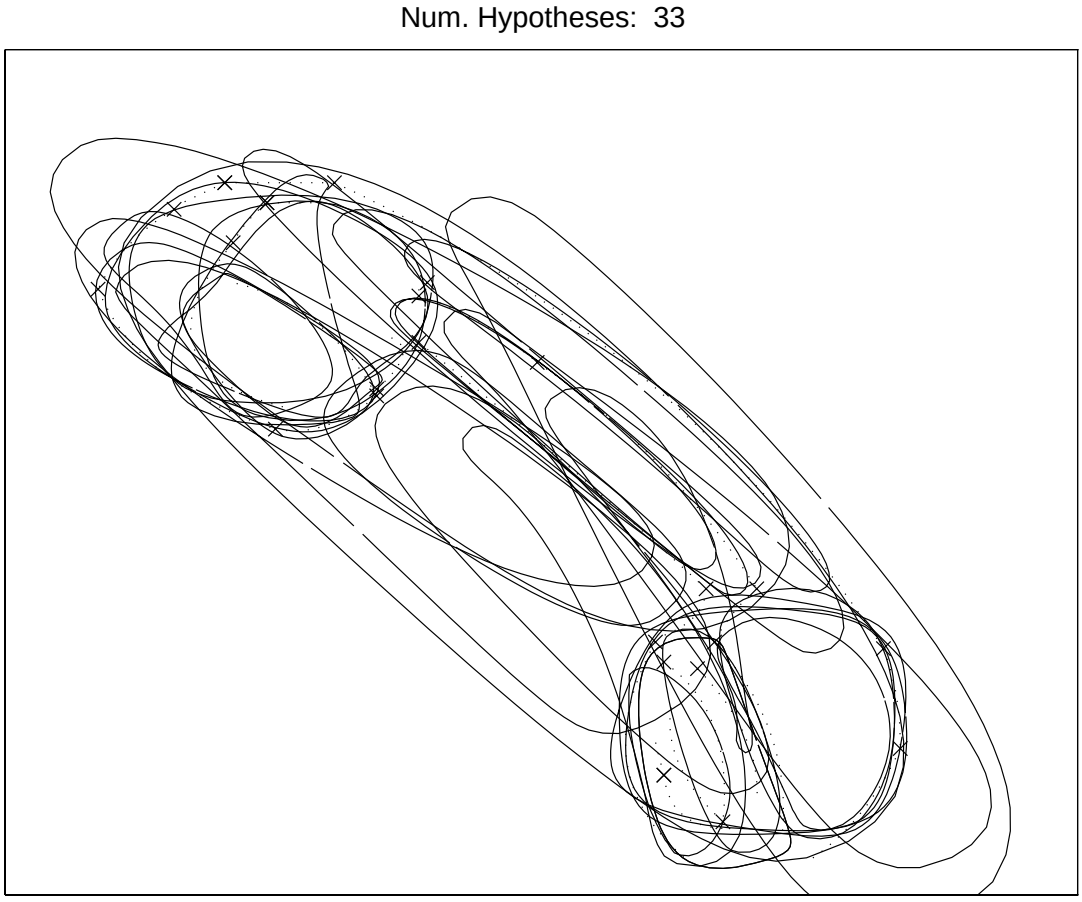


Figure 5.8: Set of part hypotheses for the handset example. The edge image and the codons can be found in Figure 4.16.

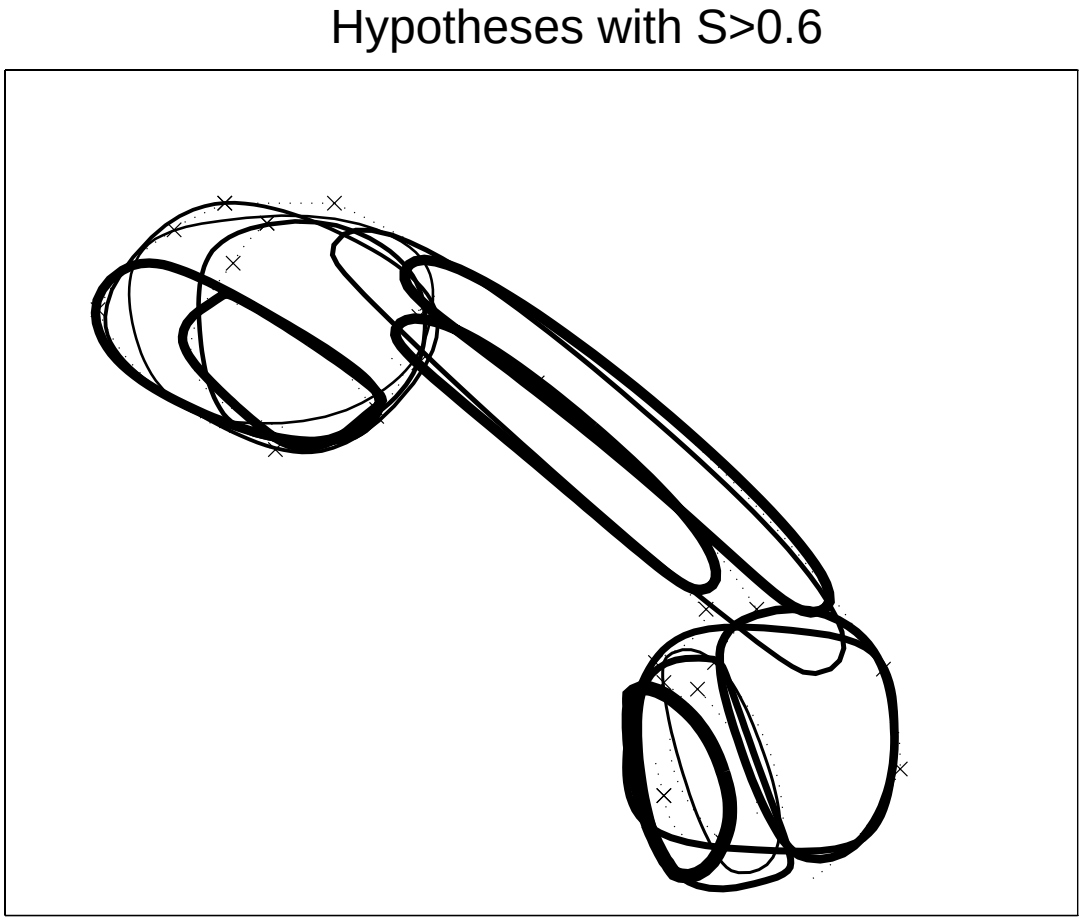


Figure 5.9: Hypotheses filtered by perceptual salience for the handset example. There are several salient groups in this case due to the pronounced tridimensionality of the image, a large shadow edge at the top and much structural detail at the bottom piece, as shown in Fig.4.16. The actual part all scored well but many hypotheses not corresponding to physical parts score the highest.

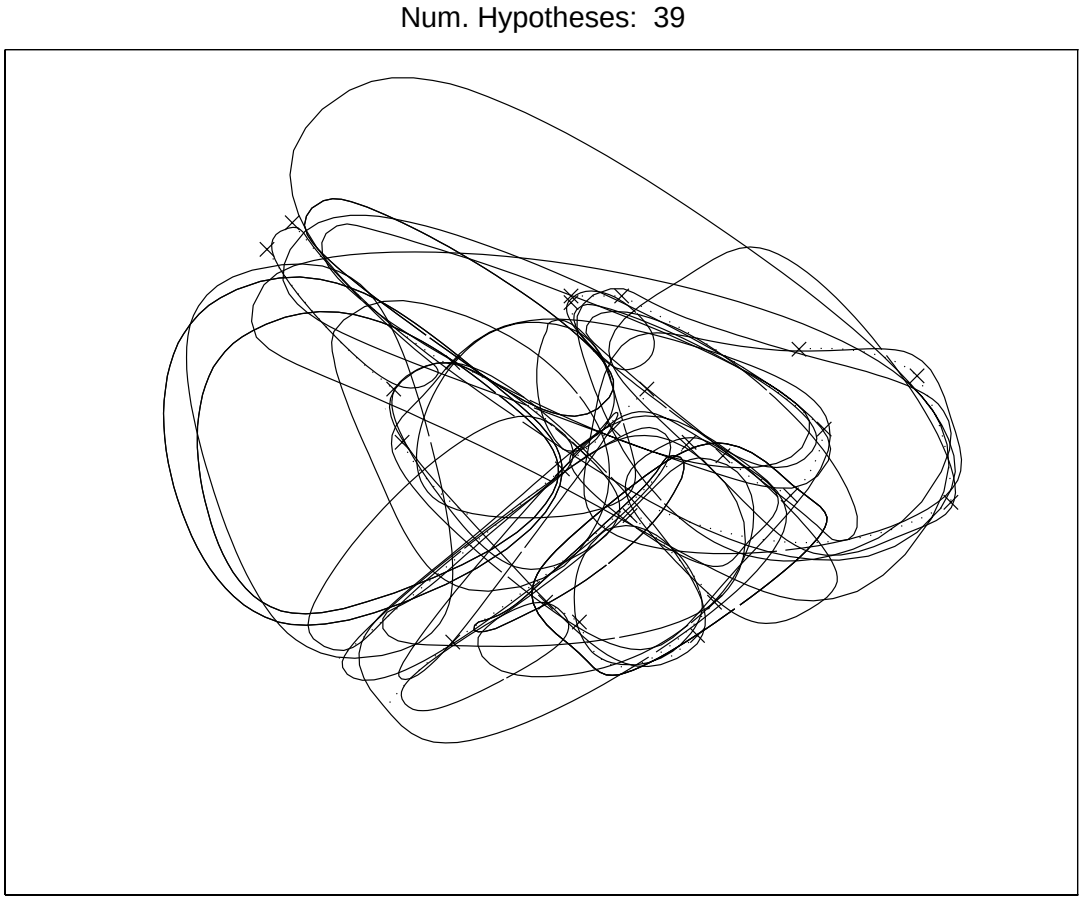


Figure 5.10: Set of part hypotheses for the beer bottle and hammer example. The edge image and the codons can be found in Figure 4.15.

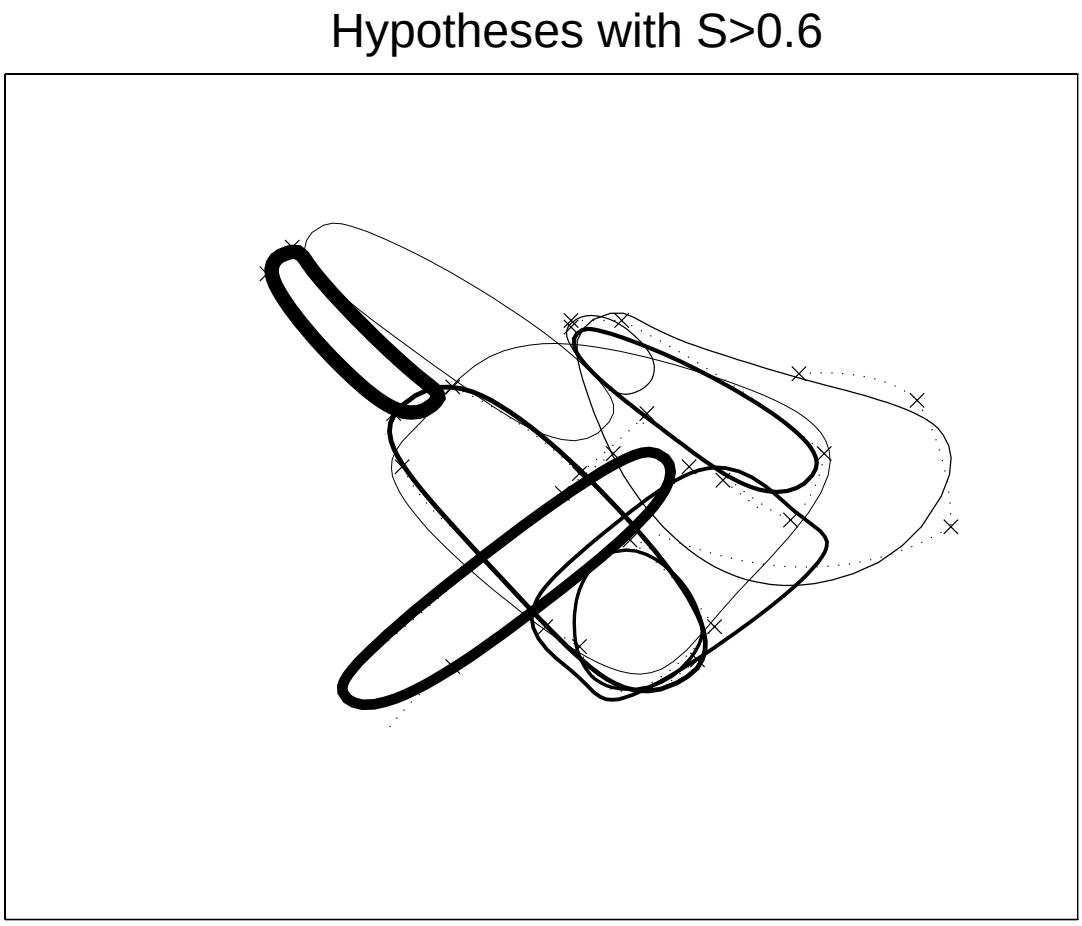


Figure 5.11: Hypotheses filtered by perceptual salience for beer bottle and hammer example. All actual parts, in particular the bottle neck, score very well apart from the occluded background object underneath the hammer head. Notice that high scores were obtained despite occlusions. Other inter-part hypotheses also have a good saliency especially the squarish one at the bottom.

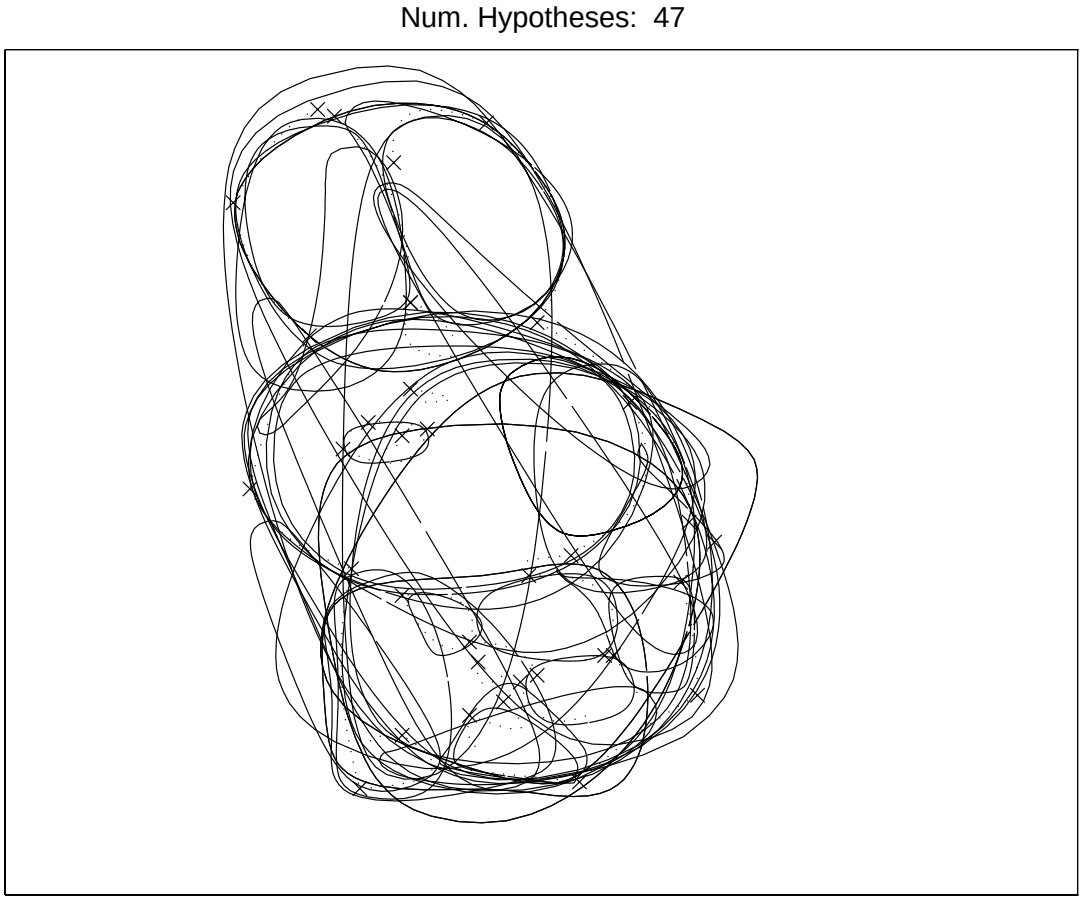


Figure 5.12: Set of part hypotheses for the rabbit example. The original intensity image and the codons can be found in Figure 4.20.

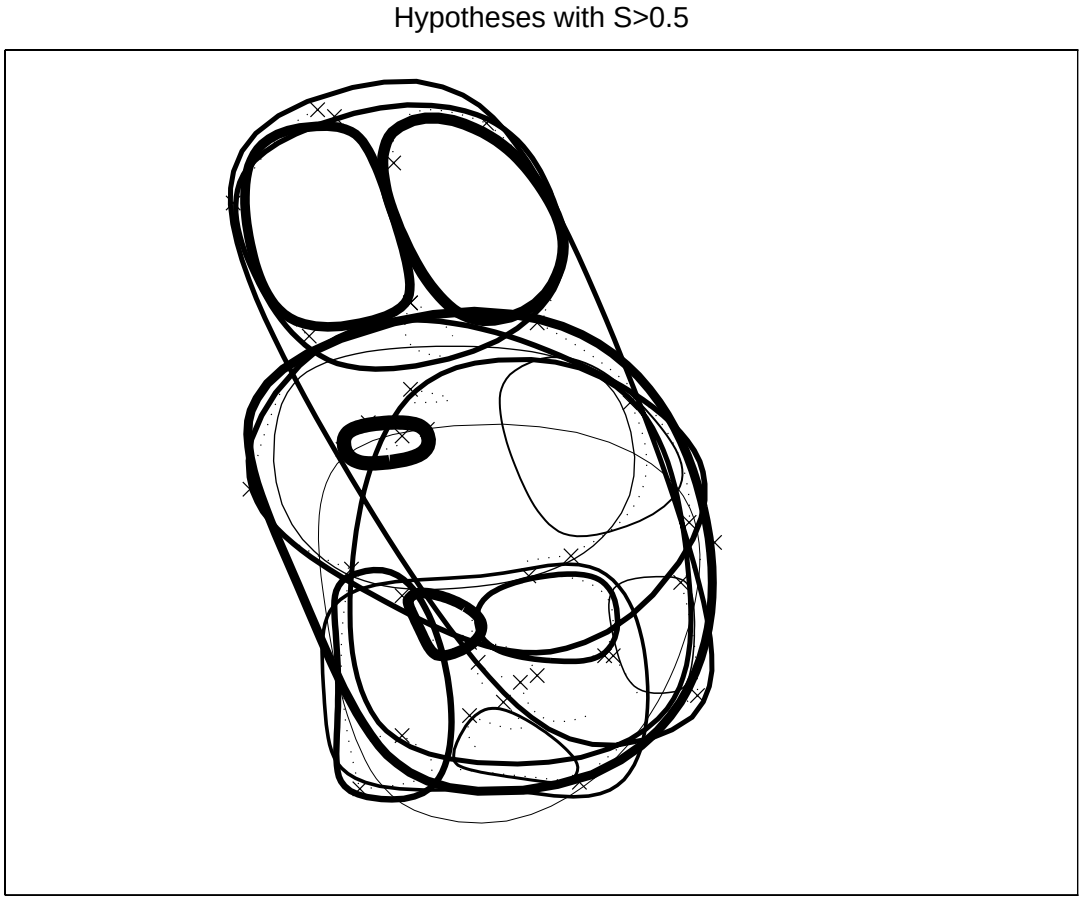


Figure 5.13: Hypotheses filtered by perceptual salience for the rabbit example with a threshold of 0.5. The nose, the two ears and a small detail below the face have the highest salience but also other actual parts, like the face and the body, score well. Many other small details have been picked that arise from some cluttering in the body that originates from the low-resolution edge image. Notably, the face has scored poorly because the top-right side of it has, unexpectedly, a codon departing from the top-right of the face and running down the left shadow which has too high displacement to be considered supportive; this drawback could be overcome by computing salience from the raw edge image instead of from the codons, as pointed out in Section 5.5.2.

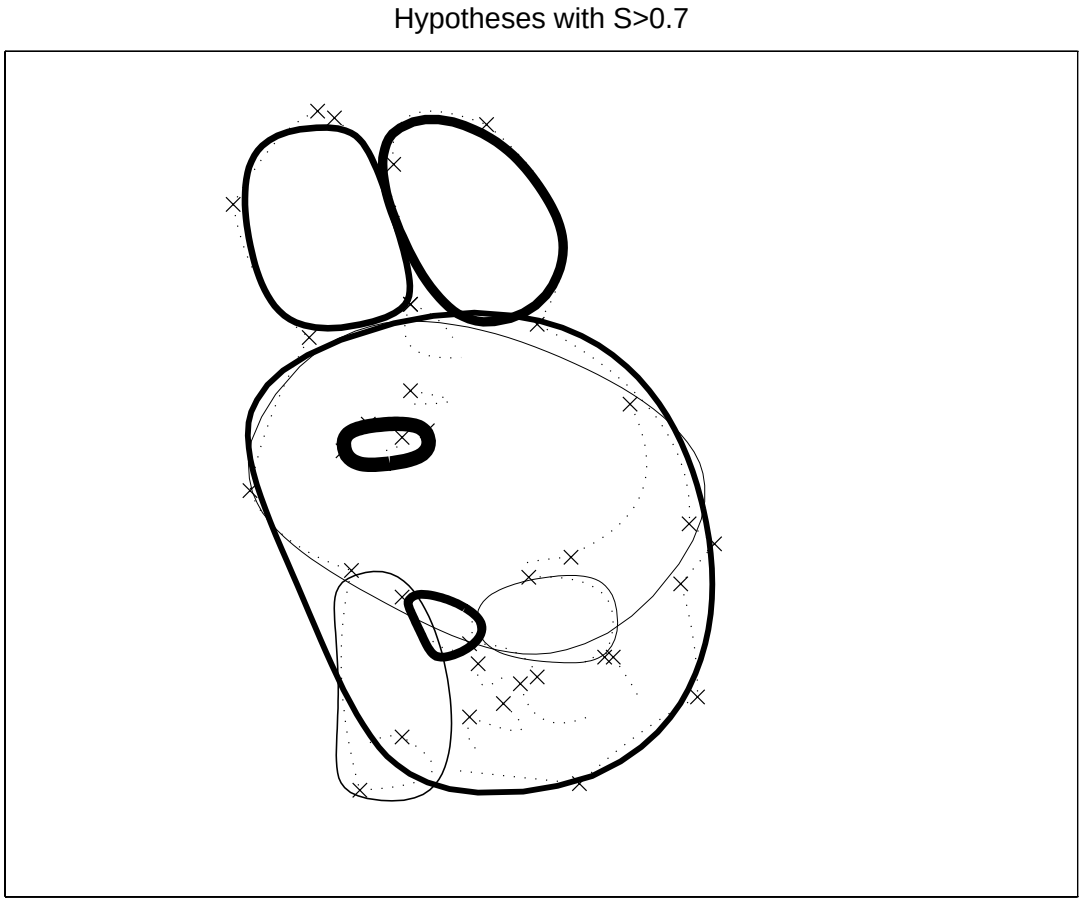


Figure 5.14: Hypotheses filtered by perceptual salience for the rabbit example with a threshold of 0.7. Referring to Figure 5.13, most hypotheses have now gone. The remaining ones are the two ears, the big hypothesis, the nose and a few spurious ones. Unfortunately, the face and the lower body have disappeared because they have low salience. However, considering the complexity of the example, the results are acceptable.

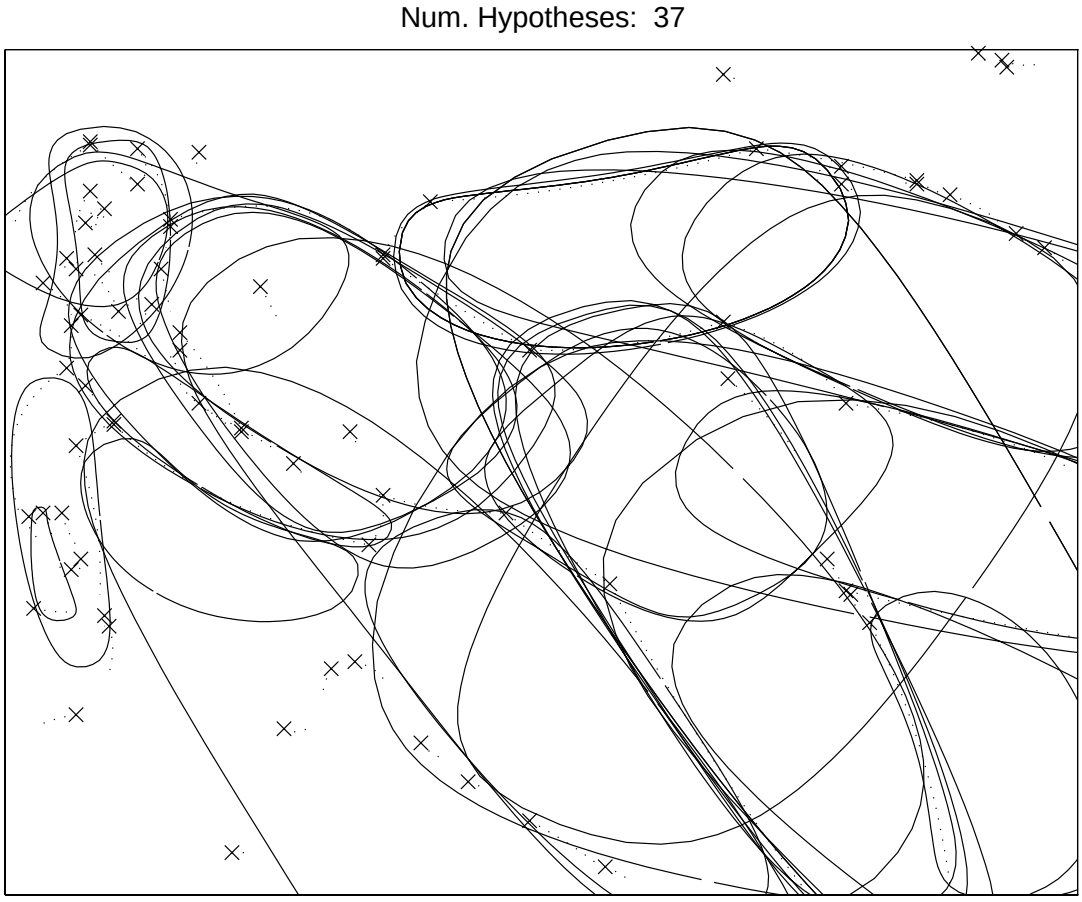


Figure 5.15: Set of part hypotheses for the Modigliani painting example. The original intensity image and the codons can be found in Figure 4.21.

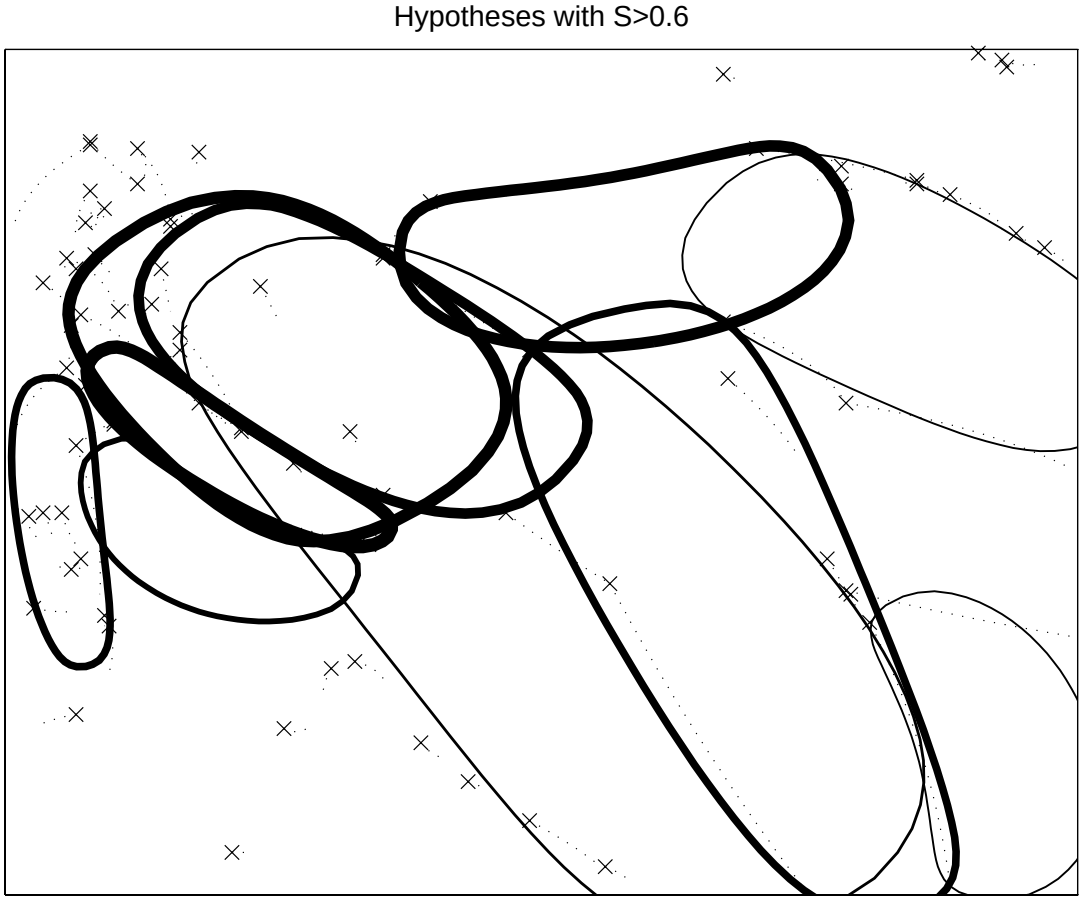


Figure 5.16: Modigliani painting example: hypotheses that have a salience greater than 0.6. This is a hard case. Only a few models score high, notably the chest, forearms and a couple of background ones. Note that the waist and the upper leg are missed because not enough edge support is available to the hypotheses.

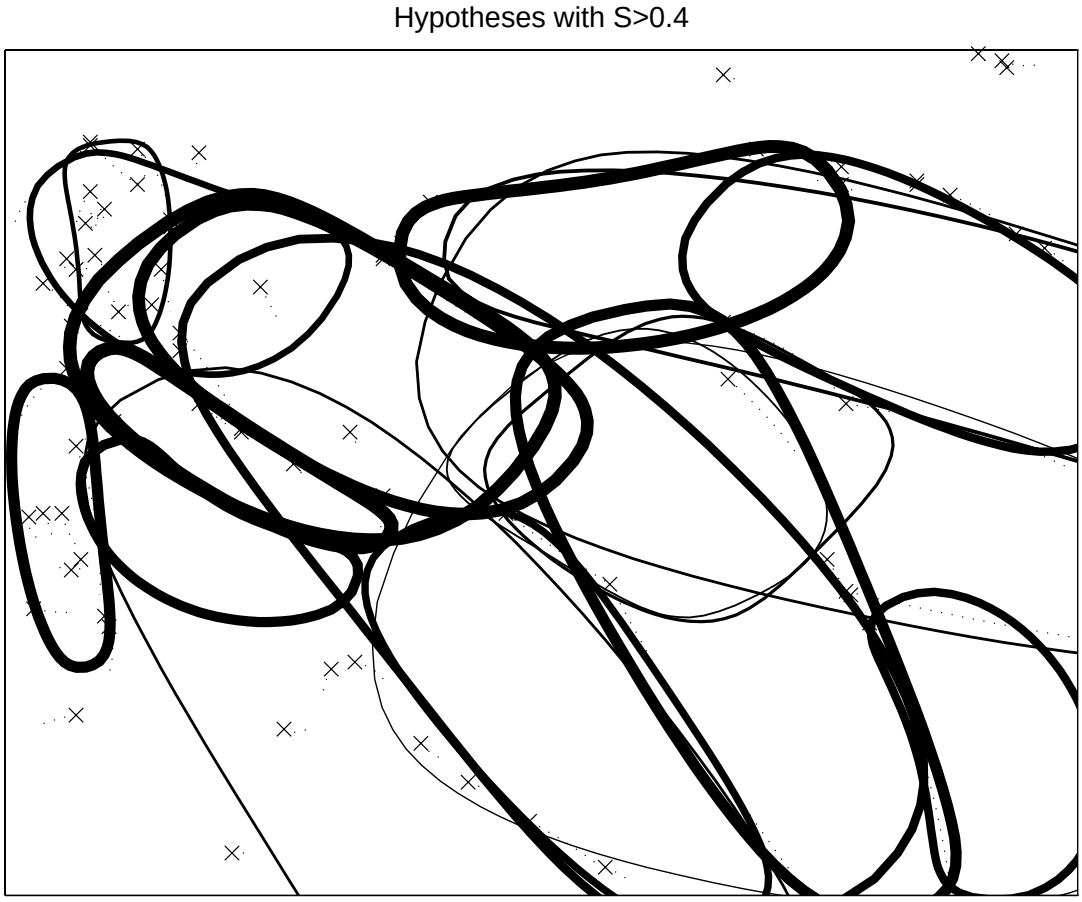


Figure 5.17: Modigliani painting example: hypotheses that have a salience greater than 0.4. All actual parts are recovered but the upper thigh (salience=0.32) is still missing for lack of local image support; in cases such as this one, strong symmetries could have been integrated to produce a more accurate representation. It can be seen that the result is rather messy because no conflict between hypotheses is accounted for; For this purpose, compare this image with the results in Figure 5.32, where the MDL filtering scheme is used.

5.3 Filtering by support competition

In this section[2] we propose a method inspired by the original works of [Pentland 90] and [Leonardis *et al.* 90] in which supporting evidence of hypotheses are put in competition with the aim of producing a minimal (or most economic) representation of the image data in terms of a few model hypotheses, hopefully corresponding to actual object parts.

Firstly, the motivation of the method and a brief review of previous work is presented, followed by an overview of the approach. Next, the Minimum Description Length cost function that accounts for supporting and conflicting evidence of the set of hypotheses and its optimisation is detailed. Many experiments are shown and some limitations of the method are discussed.

5.3.1 Motivation and related work

Let us take a look back at Figure 4.3. The whole procedure of producing the final interpretation in terms of models representing trunk and foliage can be recast into an estimation framework: we somehow use a technique that "fits" each of those models to the right data regardless disturbances caused by noise, cluttering and other entities in the image. In statistics such a technique would be termed *robust estimation* [Huber 81], which in general refers to estimation methods with outlier rejection.

A substantial difference from the well known standard robust estimation paradigm is that computer vision segmentation solutions must not only reject outliers but also deal with distinct and multiple processes, stemming from different objects, parts, surfaces and so forth.

Key to many robust estimation methods is the notion of *support regions* (or *maps*) which defines the data deemed to be originated by a single process.

Although many works make an implicit use of different support regions, e.g. the thin-plate model in [Blake & Zisserman 87], three roughly concomitant seminal works proposed the explicit use of support regions in computer vision; these are by [Pednault 89],

[2] A shorter version of this section appears in [Pilu & Fisher 96b].

[Leclerc 89] and [Pentland 90].

The introduction of the concept of support regions in computer vision allows multiple processes to be naturally dealt with. Support regions can also be disconnected and hence, in principle, occlusions could be handled in a rather unified way.

Pednault showed, by considering both residuals and supports, how curves could be segmented into polynomial patches for reconstruction purposes.

Perhaps building upon Pednault's work, Leclerc proposed and formalised an elegant framework for segmenting intensity images into regions represented by quadric patches.

Pentland generated many part hypotheses (by template matching) from silhouette which were then filtered out by explicitly taking into account extension and overlapping of hypotheses' supports and mismatches. This work was soon followed by a formalisation into a M-estimation framework by [Leonardis *et al.* 90]. His "select and recovery" iterative strategy proposed a new method for simultaneously fitting and segmenting of curves or surfaces into patches. Again [Leonardis *et al.* 94] endeavoured to use the same strategy for segmenting superquadric part models from range data, with promising results. At the same time [Darrell & Pentland 95] also developed upon the original Pentland work and produced similar results to [Leonardis *et al.* 94].

The unifying idea behind all these works is that a number of concurrent hypotheses are weighed against each other, and accepted or rejected in order to produce an "economic" representation of the image based on Occam's razor (simplicity criterion). For this purpose, they all made use of information theoretical arguments under the umbrella of the Minimum Description Length (MDL) framework [Rissnanen 83] which turned out ideal for explicitly dealing with multiple and competing hypotheses.

An excellent variation of the hypothesis competition framework that, however, does not employ MDL arguments is the one presented in [Mohan & Nevatia 92], which has already been cited in others occasions in this thesis. Their method performs filtering of the myriad of symmetry axes computed from edge images by maximising a cost function that is the sum of two terms: *i*) a weighted sum of several perceptual salience measures (such as percentage of axis covering, skew, aspect ratio and others) as a positive supporting evidence for each hypothesis; and *ii*) a not well explained conflict

measure between hypotheses which has to be negative. At first glance, it might seem that their cost function is just a rephrasing of the MDL argument – that will be described in detail in the next section – but their approach is of a more heuristic nature.

The use of the MDL criterion can be restated in Maximum Likelihood terms [Leclerc 89, Rissnanen 83], and therefore these approaches, like ours, can be properly termed as particular instances of M-estimation techniques [Huber 81]. However, the use of distinct support regions inevitably turns the segmentation into a more global problem: no matter how generated, hypotheses have to be cross-checked in a global fashion. Often this problem is formulated by embedding all the "pros and cons" of each hypothesis in a single global cost function over the whole image data which is in turn maximised (or minimised) by several means.

Thus far, the support competition method has been very promisingly used in the context of surface segmentation into quadratic patches and, as in [Leonardis *et al.* 94], to achieve part segmentation from range data. Applied to our two-dimensional segmentation problem, the method is pushed to the limits in that it has to cope with incomplete data, coarse models and multiple objects.

As in [Pentland 90] and [Leonardis *et al.* 95], *the MDL principle is used here only for selecting between competing hypotheses* that have been recovered by means of the methods discussed in the previous chapter. A different approach was followed in [Pednault 89] and [Leclerc 89] where the MDL criterion was used for both estimating and selecting. A number of experiments that I have conducted in curve fitting and segmentation by the MDL principle have shown that unless the distribution of errors is ideally Gaussian, poor results are obtained, probably due to the difficulty in estimating prior probabilities for discrete distributions, real numbers and the like. These difficulties were rightly recognised by the authors themselves and even earlier in [Witkin & Tenenbaum 85].

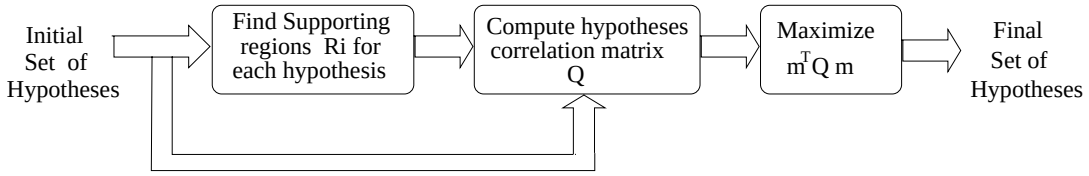


Figure 5.18: Outline of the hypothesis filtering method. From the initial set of hypotheses, supports are found and the hypotheses correlation matrix is built that accounts for supporting and conflicting evidence for all pairs of hypotheses. Then, a quadratic boolean cost function that expresses the simplicity of the solution, in the Minimum Description Length sense, is maximised with respect to the set of hypotheses $\mathbf{m}$.

5.3.2 Description of the approach

The method is related to the original work of [Pentland 90] and inspired by [Leonardis *et al.* 94], in which a number of initial estimates of superquadric models are computed by growing seed 3D patches and then filtered by optimisation of an MDL cost function. In the latter work, however, each hypothesis was a correct one, and the filtering was chiefly concerned with eliminating multiple similar hypotheses.

Although in our case many of the hypotheses generated by the method described in the previous chapter are meaningless, we use essentially the same technique, which is depicted in Figure 5.18.

Firstly, for each model $\mathcal{H}_i$ of the hypotheses set $\mathcal{H} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, a supporting region $\mathcal{R}_i$ is found by the method of Section 4.6.3 that comprises all of the codons that agree with the model.

Secondly, a matrix $\mathbf{Q}$, which we call the *hypotheses correlation matrix*, is built that takes into account interactions between hypotheses and their quality of fit. The diagonal elements $q_{i,i}$ express the goodness of fit to the supporting set of codons $\mathcal{R}_i$ of single hypotheses $\mathcal{H}_i$, whereas the off-diagonal elements $q_{i,j}$ express the interaction between the models $\mathcal{H}_i$ and $\mathcal{H}_j$ in terms of how much their support regions $\mathcal{R}_i$ and $\mathcal{R}_j$ overlap.

Let the vector $\mathbf{m} = [m_1 \; m_2 \; \cdots \; m_M]^T$ be the *hypotheses presence vector*, in which each m_i is a boolean *presence variable*, taking value "1" and "0" indicating presence or absence of the hypotheses $\mathcal{H}_i$ in the *final* description.

Then, the matrix product $\mathbf{m}^T\mathbf{Q}\mathbf{m}$ that globally expresses the simplicity of the image interpretation by the set of models $\mathbf{m}$ is maximised with respect to $\mathbf{m}$ to find a small set of models that have the highest goodness of fit to the image evidence and the least interaction between them. Multiple hypotheses are pruned because they have high correlation with others and bad hypotheses are eliminated because they do not represent image evidence as well as other hypotheses do.

One of the main advantages of this competing framework is that it does not require that model hypotheses be good in an absolute sense but how they relate to others: coarse as they might be, the best ones will surface out of this optimisation stage.

In the next subsections the construction of the matrix $\mathbf{Q}$ and its optimisation are detailed.

5.3.3 The MDL-based cost function

In the previous subsection, the basics of the simplicity principle and its relative mathematical formalism, the Minimum Description Length criterion, were briefly discussed and suggested as a guideline for filtering the large number of part hypotheses produced by the part-based grouping method of the previous chapter. In particular, it was advanced that the hypotheses filtering is performed by maximising a global quadratic boolean cost function of the form $\mathbf{m}^T\mathbf{Q}\mathbf{m}$.

In the following, it is explained what this exactly means and how $\mathbf{Q}$ is built.

Notation

Let us first introduce the notation that is going to be used to describe the MDL based cost function.

$\mathcal{E}$: the edge image; $\mathcal{E}$ has the same shape as the the original image $\mathcal{I}$ and $(i,j) \in \mathcal{E}$ is 1 if an edge has been detected at $(i,j) \in \mathcal{I}$ and 0 otherwise;

$\mathcal{C}$: the set of N codons $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$, which are the indivisible entities by which the original edge image $\mathcal{E}$ is expressed at this stage; each C_i is a connected chain of edgepoints (i,j);

$\mathcal{B}$: the set of background (non-edge) pixels; $\mathcal{B} \subseteq \mathcal{E}$ and $\mathcal{E} = \mathcal{B} + \mathcal{C}$;

$\mathcal{H}$: the set of M model hypotheses $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_M\}$ produced as in Chapter 4;

$\mathcal{X}$: a set of model hypotheses $\mathcal{X} \subseteq \mathcal{H}$

$\mathcal{R}_i$: the set of supporting codons $\mathcal{R}_i = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of a model hypothesis $\mathcal{H}_i \in \mathcal{H}$. Supporting codons are found by thresholding a proper distance norm to the model contour as defined in Section 4.6.3;

$\mathcal{R}_\mathcal{X}$: the set of support regions $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_h\}$ of a set $\mathcal{X}$ of h model hypotheses $\mathcal{X} \subset \mathcal{H}$ as defined in Section 4.6.3;

$\mathcal{B}_i$: the set of *unsupported* pixels covered by the contour of a model $\mathcal{H}_i$, in the sense illustrated in Fig. 5.1;

$\mathcal{B}_\mathcal{X}$: the set of *unsupported* pixels that are covered by the contours of the set of hypotheses $\mathcal{X}$. $\mathcal{B}_\mathcal{X} = \bigcup_{\mathcal{H}_i \in \mathcal{X}} \mathcal{B}_i$;

$\mathcal{M}_i$: the set of *supported* pixels of the hypothesis $\mathcal{H}_i$ the sense given in Section 5.2.1, that is $\mathcal{M}_i = \mathcal{R}_i \dashv \mathcal{H}_i$ (see note at Page 104);

$\mathcal{M}_\mathcal{X}$: the set of *supported* pixels of the set of hypotheses $\mathcal{X}$, that is $\mathcal{M}_\mathcal{X} = \mathcal{R} \dashv \mathcal{X}$ or, equivalently, $\mathcal{M}_\mathcal{X} = \bigcup_{\mathcal{H}_i \in \mathcal{X}} \mathcal{M}_i$

$\mathcal{M}_{i,j}$: the set of pixels of the hypothesis $\mathcal{H}_i$ (or equivalently $\mathcal{H}_j$) that are supported by the both codons $\mathcal{R}_i$ and $\mathcal{R}_j$, that is $\mathcal{M}_{i,j} = (\mathcal{R}_i \dashv \mathcal{H}_i) \cap (\mathcal{R}_j \dashv \mathcal{H}_i)$

$\mu^2(\mathcal{M}_i, \mathcal{C}_j)$: the error of fit function which expresses the displacement between the supported pixels $\mathcal{M}_i$ of a model $\mathcal{H}_i$ and one of its supporting codons $\mathcal{C}_j \in \mathcal{R}_i$. By indicating with $d(h_k, \mathcal{C}_j)$ the geometric distance of a model pixel $h_k \in \mathcal{M}_i$, to a codon $\mathcal{C}_j$, the error of fit function is defined as $\mu^2(\mathcal{M}_i, \mathcal{C}_j) = \sum_{h_k \in \mathcal{M}_i} d(h_k, \mathcal{C}_j)^2$

$\mu^2(\mathcal{M}_\mathcal{X}, \mathcal{R}_\mathcal{X})$: the error of fit function which expresses the displacement between the supported pixels $\mathcal{M}_\mathcal{X}$ of set of models $\mathcal{X} \subseteq \mathcal{H}$ and its supporting codons $\mathcal{R}_\mathcal{X}$. $\mu^2(\mathcal{M}_\mathcal{X}, \mathcal{R}_\mathcal{X}) = \sum_{\mathcal{M}_i \in \mathcal{M}_\mathcal{X}} \sum_{\mathcal{C}_j \in \mathcal{R}_i} \mu^2(\mathcal{M}_i, \mathcal{C}_j)$.

Rationale

Let us now explicitly formulate the problem in MDL terms. Let us indicate by $L(\cdot)$ a generic function that gives the number of bits needed to represent a certain entity. The precise definition of L with respect to a certain entity will be given later.

Since the edge image $\mathcal{E}$ can be decomposed into two distinct elements, namely the background and the codons, the number of bits needed to represent it can be written as:

$$L(\mathcal{E}) = L(\mathcal{C}) + L(\mathcal{B})$$

When we interpret part of the edge image $\mathcal{E}$ by a set of models $\mathcal{X}$, the encoding length changes to what we indicate with $L(\mathcal{E}|\mathcal{X})$[3]:

$$L(\mathcal{E}|\mathcal{X}) = L(\mathcal{E}) - L(\mathcal{M}_{\mathcal{X}}) + L(\mathcal{B}_{\mathcal{X}}) + L(\mu^2(\mathcal{M}_{\mathcal{X}}, \mathcal{R}_{\mathcal{X}})) + L(\mathcal{X}) \tag{5.3}$$

There are four new terms that contribute to changing the original encoding length:

$-L(\mathcal{M}_{\mathcal{X}})$**:** this negative term represent the saving due to support regions being now described by supported portions $\mathcal{M}_{\mathcal{X}}$ of the set of models $\mathcal{X}$; contours of the set of models $\mathcal{X}$;

$+L(\mu^2(\mathcal{M}_{\mathcal{X}}, \mathcal{R}_{\mathcal{X}}))$**:** this expresses the additional number of bits needed to express (somehow) the displacement between support regions and supported model contours;

$+L(\mathcal{B}_{\mathcal{X}})$**:** this positive term represent the additional cost of having to express unsupported portions of the contour model. It rarely occurs that a model is fully supported along its contour and therefore the cost of specifying "gaps" has also to be taken into account, differently from [Leonardis *et al.* 95], where simply connected support regions were assumed. This term is of particular importance and will be discussed upon later.

[3] The symbol "|" resembles the conditional probability notation and it was used in the same context by [Leclerc 89].

$+L(\mathcal{X})$: this positive term, called model overhead, is the additional cost of having to express the parameters of the models;

In the MDL framework, the search for the minimal subset of models $\hat{\mathcal{H}}$ of $\mathcal{H}$ that gives an optimal description of the set codons $\mathcal{C}$ if performed by finding a subset $\mathcal{X} = \hat{\mathcal{H}} \subseteq \mathcal{H}$ that minimises the encoding length $L(\mathcal{E}|\mathcal{X})$. In formal terms:

$$\hat{\mathcal{H}} = \arg\min_{\mathcal{X} \subseteq \mathcal{H}} \{L(\mathcal{E}|\mathcal{X})\}$$

By using the definition given in Eqn. (5.3) and by noticing that the term $L(\mathcal{E})$ is a constant, the above minimisation becomes:

$$\hat{\mathcal{H}} = \arg\max_{\mathcal{X} \subseteq \mathcal{H}} \Big\{ L(\mathcal{M}_{\mathcal{X}}) - L(\mathcal{B}_{\mathcal{X}}) - L(\mu^2(\mathcal{X}, \mathcal{R}_{\mathcal{X}})) - L(\mathcal{X}) \Big\} \tag{5.4}$$

The maximiser expression in braces, which we call $S(\mathcal{E}|\mathcal{X})$, is normally termed as "bit saving", because in fact it represents the decrease of encoding length due to the use of models.

Although the terms representing the model overhead and residual encoding have always a negative sign (which pulls down the overall bit saving), a few words must be spent on the signs of $L(\mathcal{M}_{\mathcal{X}})$ and $L(\mathcal{B}_{\mathcal{X}})$.

As we know from classical information theory, the minimum number of bits (in the Hoffman sense) needed to encode the data generated by a stochastic process – as an edge image can in general be considered – equals the negative base-two logarithm of the probability of observing that data [Leclerc 89]. Normally, in an edge image the probability of having an edge at a given location is much lower than being a non-edge and therefore edge pixels have longer average encoding than background pixels (e.g., the paper on a fax machine speeds up when receiving background). As a result, when portions of edge data are represented by a compact model there is a considerable bit saving; on the other hand, when the background is to be specified there is actually an overhead. From these intuitive considerations follows their respective positive and negative contribution to the overall bit saving.

In Section 5.3.4 a formal account of these two contribution is given but in the following no prior knowledge of edge and background pixels occurrence probability is assumed other than that reflected in their contribution sign to the overall encoding length function.

Practical formulation

Let us now suppose we can determine four constants K_1, K_2, K_3 and K_4 such that:

K_1 is the average number of bits necessary to represent each *supported* pixel of a model contour;

K_2 is the average number of bits necessary to represent each *unsupported* pixel of a model contour;

K_3 is a constant such that when multiplied by $\mu^2(\mathcal{X}, \mathcal{C})$ gives the average encoding length for representing the residuals;

K_4 is the average number of bits for specifying the parameters of a model.

Then, following the philosophy of [Leonardis *et al.* 95], we can rewrite the bit saving $S(\mathcal{E}|\mathcal{X})$ as follows:

$$S(\mathcal{E}|\mathcal{X}) = \overbrace{K_1 \cdot |\mathcal{M}_\mathcal{X}|}^{\mathbf{a}} - \overbrace{K_2 \cdot |\mathcal{B}_\mathcal{X}|}^{\mathbf{b}} - \overbrace{K_3 \cdot \mu^2(\mathcal{M}_\mathcal{X}, \mathcal{R}_\mathcal{X})}^{\mathbf{c}} - \overbrace{\sum_{\mathcal{H}_i \in \mathcal{X}} K_4}^{\mathbf{d}}. \qquad (5.5)$$

where $|\cdot|$ indicates the number of image pixels represented by $\mathcal{M}_\mathcal{X}$ and $\mathcal{B}_\mathcal{X}$, respectively. As previously said, the terms under braces account for:

a : the number of bits saved by expressing supported codons by the models; it is fundamental that savings due to supports and residual overheads are not accounted more than once when portions of contour are shared by the same models in the final description [Pentland 90].

b : the additional cost in bits of having to express *unsupported* parts of the models;

c : the cost in bits of expressing the displacement between model and supporting codons;

d : the cost in bits for specifying the parameters of all the models $\mathcal{H}_i \in \mathcal{X}$.

The inclusion of the term **b** gives favour to models which have higher contour covering and constitute a fundamental variation with respect to the MDL cost functions used in [Darrell & Pentland 95] and [Leonardis *et al.* 95]. Without this term, models could be selected regardless the amount of unsupported contour portions, often leading to solutions such as the one shown in Figure 5.26-B.

Algorithmic formulation

If we assume that in the *final* solution the only kind of model overlapping taken into account is pairwise[4] [Pentland 90], the maximisation in Eqn. (5.4) can be achieved by transforming Equation (5.5) into a more compact matrix form, which is derived from [Leonardis *et al.* 95]. This pairwise overlapping assumption is a fairly sensible choice that helps keep the computational cost down, eases optimisation and is justified by the fact that three or more parts are very seldom jointly together in the same region.

Under this assumption, the maximisation can be rewritten as:

$$\hat{\mathbf{m}} = \arg\max_{\mathbf{m}} \left\{ \mathbf{m}^T \mathbf{Q} \mathbf{m} \right\} \tag{5.6}$$

where $\mathbf{Q}$ is the hypotheses correlation matrix, which will be defined next, and $\mathbf{m} = [m_1 \; m_2 \; \cdots \; m_M]^T$ is the hypotheses presence vector in which each element m_i is "1" or "0" if the model $\mathcal{H}_i$ is present or absent, respectively, in the final image description; any given $\mathbf{m}$ selects a subset $\mathcal{X}$ of the whole set of hypotheses $\mathcal{H}$.

Each diagonal element $q_{i,i}$ of $\mathbf{Q}$ expresses the length of encoding the supporting region $\mathcal{R}_i$ of a hypothesis $\mathcal{H}_i$ by $\mathcal{H}_i$ itself:

$$q_{i,i} = K_1|\mathcal{M}_i| - K_2|\mathcal{B}_i| - K_3\mu^2(\mathcal{M}_i, \mathcal{R}_i) - K_4; \tag{5.7}$$

[4] Differently from [Pentland 90], overlapping here refers to sharing codons.

The off-diagonal elements $q_{i,j}$ deal with interaction between two competing (possibly partially overlapping) hypotheses $\mathcal{H}_i$ and $\mathcal{H}_j$:

$$q_{j,i} = q_{i,j} = \frac{1}{2}\left\{ -\overbrace{K_1 \cdot |\mathcal{M}_{i,j}|}^{\mathbf{e}} + \overbrace{K_3 \cdot \mu^2(\mathcal{M}_{i,j}, \mathcal{R}_i \cap \mathcal{R}_j)}^{\mathbf{f}} \right\} \tag{5.8}$$

The term **e** expresses the number of image pixels that are supported by both model $\mathcal{H}_i$ and $\mathcal{H}_j$ and since two models rarely overlap it can be approximated by taking the number of pixels of just $\mathcal{H}_i$ that are supported by the shared codons. The term indicated by **f** carries the cost of expressing the residuals for codons that are shared by both $\mathcal{H}_i$ and $\mathcal{H}_j$. The off-diagonal terms ensure that saving and residual overhead due to shared supports are accounted for only once.

Intuitively, with this definition, $\mathbf{m}^T\mathbf{Q}\mathbf{m}$ is large when the smallest number of models best describe the image and do not have too many unsupported contour portions.

In a later subsection, the optimisation procedure will be discussed along with a brief literature review of methods used in similar problems. The next section will discuss how the constants K_1, K_2, K_3 and K_4 are qualitatively determined.

5.3.4 On the determination of the constants

The MDL principle states that the choice of the constants K_1, K_2, K_3 and K_4 should be theoretically guided by prior probability distributions of edges, gaps, residual and model parameters.

The determination of the two constants K_1 and K_2 is a very challenging task. An interesting attempt to estimate something similar to K_1 and K_2 in a more formal context is given in Section 6.5.1 (see also [Pilu & Fisher 96d]). If p_{m1} is the probability that a pixel on a model contour is supported (i.e. matching a feature) and if p_{b1} is the probability of detecting an edge at a certain image pixel, then by comparing Eqn. (6.8) and Eqn. (5.5) we have:

$$\begin{aligned} K_1 &= \log_2(p_{m1}) - \log_2(p_{b1}) \\ K_2 &= -(\ \log_2(1-p_{m1}) + log_2(1-p_{b1})\ ) \end{aligned} \tag{5.9}$$

For instance, for $p_{m1}=0.8$ (i.e. 80% of the model contour is expected to be supported by codon evidence) and $p_{b1}=0.05$ (i.e. 5% of the image pixels are expected to be edges) we obtain $K_1=4$ and $K_2=2.3$, which are amazingly close to what the experiments (Sec. 5.3.6) indicated as an optimal combination. In Section 5.3.3 it was argued that the contribution of $L(\mathcal{B}_{\mathcal{X}})$ to the bit saving was negative and, as a matter of fact, the negative sign of K_2 in Eqn. (5.9) suggests that it is so for real-case values of p_{m1} and p_{b1}!

The error of fit (EoF) $\mu^2(\mathcal{M}_i, \mathcal{C}_j)$ is used also in [Leonardis *et al.* 95]. If the distribution of model/codon displacements is Gaussian with variance $\sigma^2_{i,j}$, it can be easily[5], demonstrated that [Leclerc 89, Leonardis *et al.* 95]:

$$K_3 \approx \frac{\log_2 \sigma_{i,j} + \frac{1}{2}\log_2 2\pi e}{\sigma^2_{i,j}} \tag{5.10}$$

The value of K_3 decreases with increasing noise level and this would suggest that also in our non-Gaussian case, K_3 should depend on the magnitude of the displacement. Since it is not clear how this should be done from even a detailed knowledge of the displacement probability distribution, the value of K_3 currently does not vary with the fitting quality. However, Eqn. (5.10) helps determine a possible range for K_3. For instance, with a typical 128x128 image (the one most commonly used for experiments in this thesis) $\sigma_{i,j}$ varies from 2 to 4 pixels and by Eqn. (5.10) K_3 would go from 0.7 to 0.25, respectively; as we shall see in in the experiments of Sec. 5.3.6, these values are a little bit too high, probably because our noise distribution is not Gaussian.

The value of K_4 represents the number of bits necessary to encode the model parameters. As argued in, e.g., [Pednault 89] and [Leclerc 89], it is very hard to determine prior probabilities for floating point numbers. In addition to that, if an optimal parameter encoding is sought, the distribution of parameters across the whole training set should be considered. In spite of all these difficulties, a good range value of K_4 has been experimentally found to be from 20 to 80, although the lower bound seems a priori a very conservative estimate.

[5] In the case of Gaussian noise it is possible to derive the average coding length needed to minimally represent (in the Huffman sense) the data; see [Leclerc 89].

Since these four constants could be arbitrarily scaled, in [Leonardis 93] it was suggested that one of them could be set to 1 while experimenting with the others; here I have preferred to keep their natural values.

Having suggested values for these constants, I do not whish to further conjecture about exotic methods for estimating prior probabilities and the reasons of this choice are also made clear in [Leonardis *et al.* 95]: no good theory is available at the moment for estimating these prior probabilities other than for the unrealistic Gaussian noise situation, and if it did, neither would it be useful, for the assumptions it would make about the problem would probably not generalise across all images.

5.3.5 Optimisation

Equation (5.6) is, technically speaking, a *quadratic boolean optimisation problem*, as the solution space can be represented as a corner of an M-dimensional hypercube.

In [Leonardis *et al.* 95] and [Pentland 90] the optimisation problem was tackled by using different strategies, mainly reflecting the kind of problem they had to deal with. Neither of the methods they used would function for our problem, which is substantially different from theirs.

Pentland had a set of hypotheses, many of which were rather poor because of the simple template matching technique used for generating them. He performed the optimisation, perhaps after [Leclerc 89], by using a continuation method. The matrix $\mathbf{Q}$ is rendered positive definite by adding a positive constant k to its diagonal. This turns $\mathbf{m}^T(\mathbf{Q}+k\mathbf{I})\mathbf{m}$ into a convex function of $\mathbf{m}$ and a solution is found by a gradient descent method. Then, the constant k is progressively reduced and, by employing gradient descent at each step, the maximum is tracked all the way up to the final solution for $k = 0$. Unfortunately, our cost function is much more ragged and with narrower peaks than Pentland's because the quality of our hypotheses is generally very poor. In fact early experiments we carried out using this technique showed that a plausible solution is never obtained except in few exceptional cases where, not surprisingly, the hypotheses were good. In [Darrell & Pentland 95] too, gradient descent was successfully used but the framework proposed there was pretty much equivalent to [Leonardis *et al.* 95].

In [Leonardis *et al.* 95] an even more straightforward method was used, a greedy winner-takes-all optimisation. His good results were due to the clear statement that the input should be "well-behaved". In fact, his major concern was to eliminate overlapping models and once one was selected, the optimisation reduced to just rejecting other overlapping models, thereby justifying the use the the winner-takes-all method.

Needless say, those strategies would not be of any use to our problem because we do not have, in general, good hypotheses; an ideal optimisation procedure employed should, more than anything else, try to find out which are best by putting them in relation to others. This problem is of an inherently global nature – in principle combinatorial – and needs a more powerful optimisation tool to be successfully solved.

An interesting approach was used in [Mohan & Nevatia 92], where the maximisation of their cost function (reducible to a quadratic form) was achieved by a constraint satisfaction network and the final solution was obtained by thresholding the outputs of the network after convergence was achieved; it might look that by doing so a solution with different confidence could be produced, thereby avoiding the in/out nature of the solution given by the boolean vector used here[6]. However, the network is clearly not a probabilistic one, given that no probabilities are encoded in the system, and such a claim is without support. Nevertheless, the use of a probabilistic network solution, such as the one used for a perceptual organisation problem in [Sakar & Boyer 93], is certainly a promising path to be explored in future work.

As far as the optimisation strategy employed here is concerned, in earlier experiments the boolean quadratic form $L(\mathbf{m}) = \mathbf{m}^T\mathbf{Q}\mathbf{m}$ was maximised by simulated annealing. Simulated annealing (which is described in Appendix C) is a powerful optimisation method that has been often used for solving boolean problems. Although the initial results looked very promising, a more careful analysis revealed that the hypotheses that were most likely to be detected were the ones that scored the highest in isolation, yielding good but still sub-optimal minima. This is not surprising, because the hill-climbing direction in the space of $\mathbf{m}$ is always biased to hypotheses that contributed most to the cost function in isolation.

[6] In an earlier paper [Mohan & Nevatia 89], a winner-takes-all type of network was superimposed on the constraint satisfaction network to prevent bad hypotheses from being selected for lack of competition with others.

Since the intention was to investigate the real properties and limitations of the proposed segmentation method in the optimal case, a simple genetic algorithm was implemented[7].

The chromosomes used are simply instances of the boolean hypotheses presence vector $\mathbf{m}$, which is a natural choice for boolean optimisation. A good population size was found to be 500. The initial population was chosen keeping in mind that it is likely that just 10-15% of the initial set of hypotheses are the correct ones and therefore a random 10 % of the genes were set to one. It has been noticed that this heuristic choice yielded quicker convergence than using a random initial population. Figure 5.19 shows two GA runs with 10% (left) and 90% (right) of the genes set to one: in the former case convergence is achieved relatively more quickly.

As far as the reproduction strategy is concerned, twins are generated by one-point crossover of two parents picked by rank-based over-selection of the 50% fittest members of the population; the parents are put in the population pool along with the generated twins to make up a new generation. By this elitist strategy, the fittest elements are more likely to breed with the fittest members. This, although believed to yield genetic drift, has led in all cases to relatively fast convergence (30-40 generations) with respect to the classical tournament selection reproduction method. The mutation probability was set to a 1% value. However, as shown, e.g., in [Grefensette 96], the basic mechanism of GA is so robust that the parameter tuning has not been found to be critical for the quality of convergence.

A few experiments have been carried out in a real case with 16 real hypotheses to see whether the maximum obtained by the GA was the global one, which was found off line by computing the cost function for all the $2^{16} = 65536$ combinations; in all cases, the solution obtained corresponded to the theoretical maximum. Of course, for larger hypothesis sets no tests are possible since, for instance, it would take 34 years to exhaustively test all the combinations of 40 hypotheses if the evaluation of the cost function took a mere 1ms!

The genetic algorithm has been implemented in Matlab and the running time is about

[7] Thanks to A.W. Fitzgibbon for the idea.

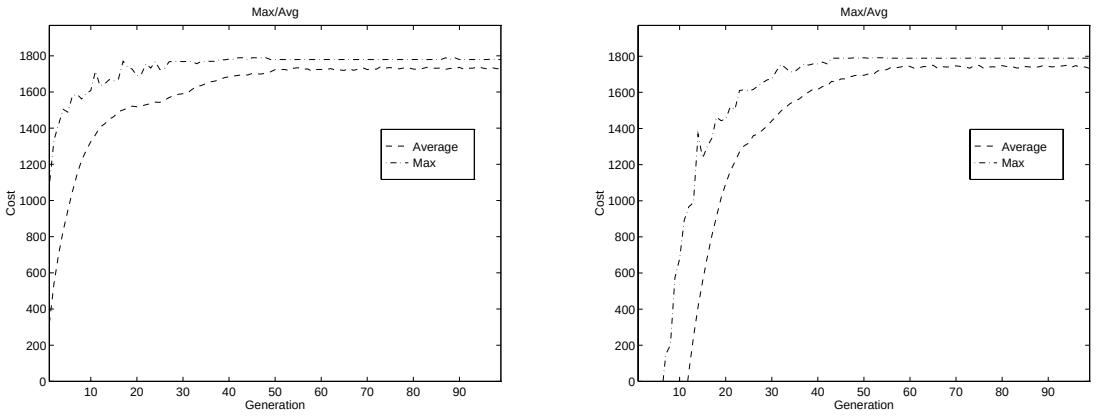


Figure 5.19: Populations for different starting points. The left figure shows a run for an initialisation with a random 10% of the genes set to one, whereas on the other figure the number of ones is 90%. Notice the faster convergence in the first case.

20s on a SPARC 10 machine for 50 generations and 50 hypotheses. As noticed by [Pentland 90], the matrix $\mathbf{Q}$ is symmetric and banded and this could be used to reduce the load of computing the cost (fitness) function $\mathbf{m}^T\mathbf{Q}\mathbf{m}$ to further speed up the whole process. I reckon that an optimised implementation of the GA would run in one tenth of the time under the same conditions.

5.3.6 Experimental results

In this section, filtering experiments by the competitive MDL method are shown for the same set of images as the one used in Section 5.2. Each experiment is described in the relative caption so here an overview and some comparison are given.

For each of the first two test images – tree in Figs. 5.20-5.22 and screw-driver in Figs. 5.23-5.25 – three experiments are given that illustrate the sensitivity of the solutions to changes in K_4, K_3 and for some combinations of K_1 and K_2. Then, some similar experiments are shown for the handset (Fig. 5.26-5.27), beer bottle (Fig. 5.28) and hand examples (Fig. 5.29-5.29) that are more oriented to help assess the goodness of solution for the same parameter configurations.

For these five examples, good solutions are always obtained with the same parameter configurations, i.e. $K_1 = 3.6$, $K_2 = 2.5$, $K_3 = 0.1$ and $K_4 = 40$, but the insensitivity to

rather large variations in the parameters is noteworthy. Moreover, when errors crop up, they do it in an equally explainable fashion for all experiments, which further supports the claim of stability of the MDL based method.

In the last two examples with the toy rabbit in Fig. 5.31 and the Modigliani painting in Fig. 5.32, less exciting results are obtained: the presence of high cluttering and unclear part separation have hindered a perfect generation of part hypotheses and a high figure-ground ambiguity (in the case of the Modigliani painting) has made filtering impossible. These problems have already been discussed in Section 5.2.2.

Finally, some limitations of the method, in particular its incapacity to discriminate between some ambiguous situations, have been noticed and will be fully explained in the next subsection.

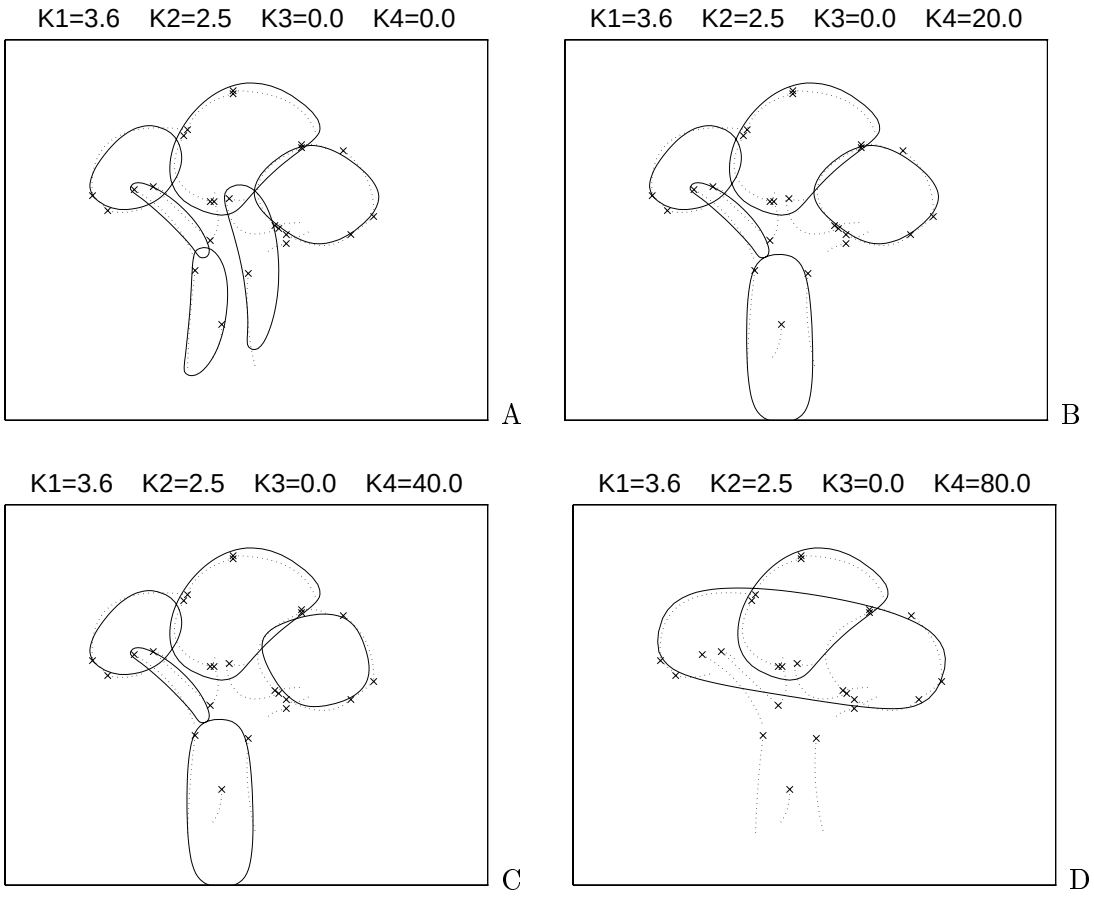


Figure 5.20: Filtering experiments by MDL for the tree example with different values of the model overhead K_4. As K_4 grows, fewer models describe the data. In particular, for $K_4 = 0$ a spurious PDM describes the right side of the trunk, whereas the small detail in the centre is taken up by another model. There is a wide range of K_4 (10 to 70) for which the result is the intuitively correct one. As pointed out in Sec. 4.7, the two branches are not recovered because they are at too small a scale. For $K_4 = 80$, not only is the trunk lost (not enough support to justify the cost of the model) but a large hypothesis crops up that embraces the two opposite bushes.

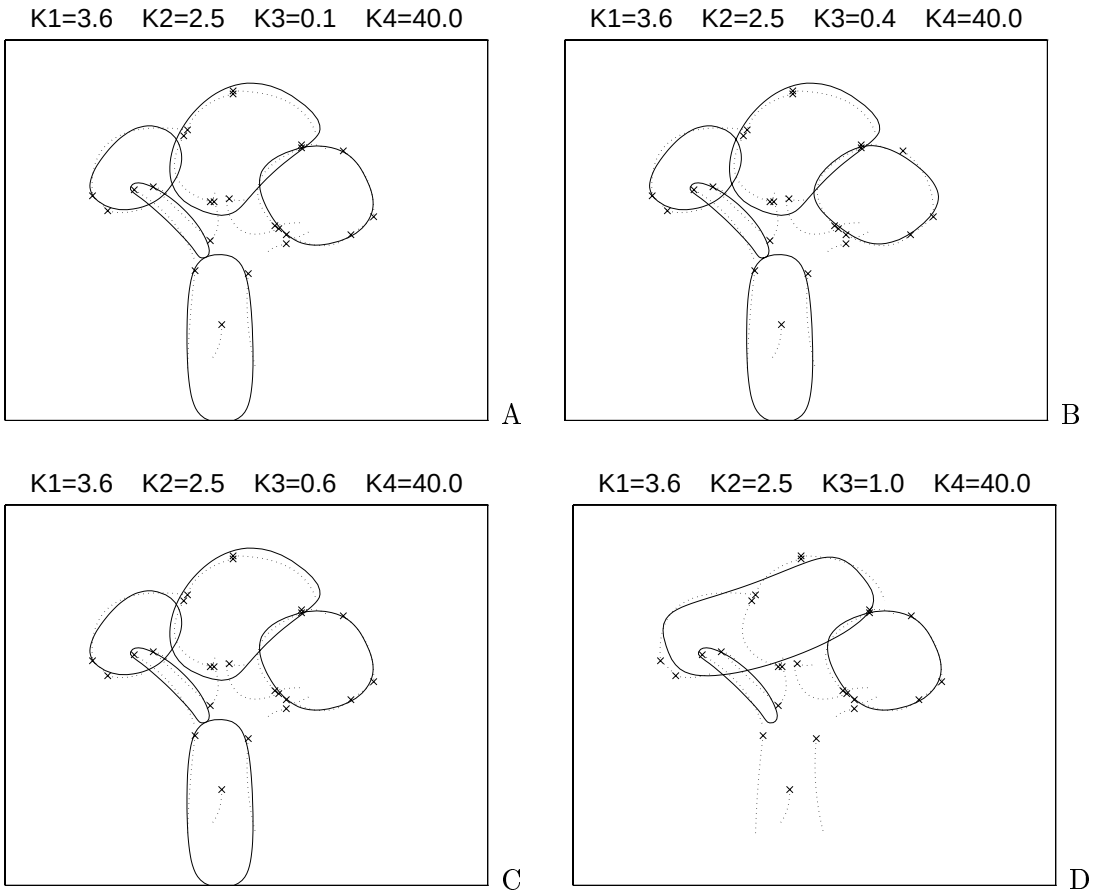


Figure 5.21: Filtering experiments by MDL for the tree example with different values of the residual cost factor K_3, keeping $K_1 = 3.6$, $K_2 = 2.5$ and $K_4 = 40$ fixed. It can be seen that the correct segmentation is achieved for a large range of K_3 (figures A, B and C), except when it gets too big, in this case greater than 0.8. In fig. D the result for $K_3 = 1.0$ is similar to what happened in Fig. 5.20-D, that is, the cost of expressing the residuals gets too high to justify the presence of too many models.

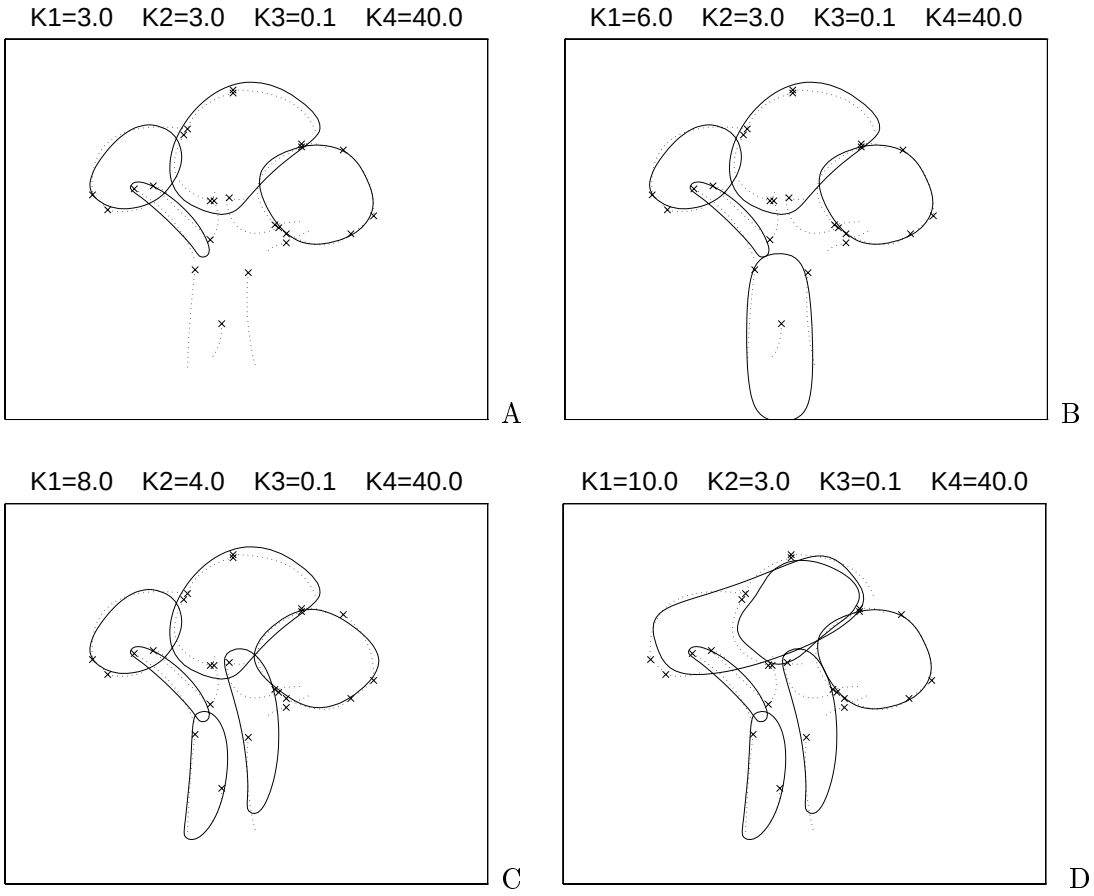


Figure 5.22: Filtering experiments by MDL for the tree example with different values of the constants K_1 and K_2, keeping $K_3 = 0.1$ and $K_4 = 40$ fixed. It is worth noticing the stable presence of the three bushes and the left branch until K_1 gets too big with respect to K_2, when a bigger model "grabs" the two bushes because of the reduced cost of expressing its unsupported lower region. The situation of the trunk is again unstable, with its actual hypothesis appearing only in fig. B; in the other cases we have the same phenomena as in Figures 5.20-A and 5.20-D.

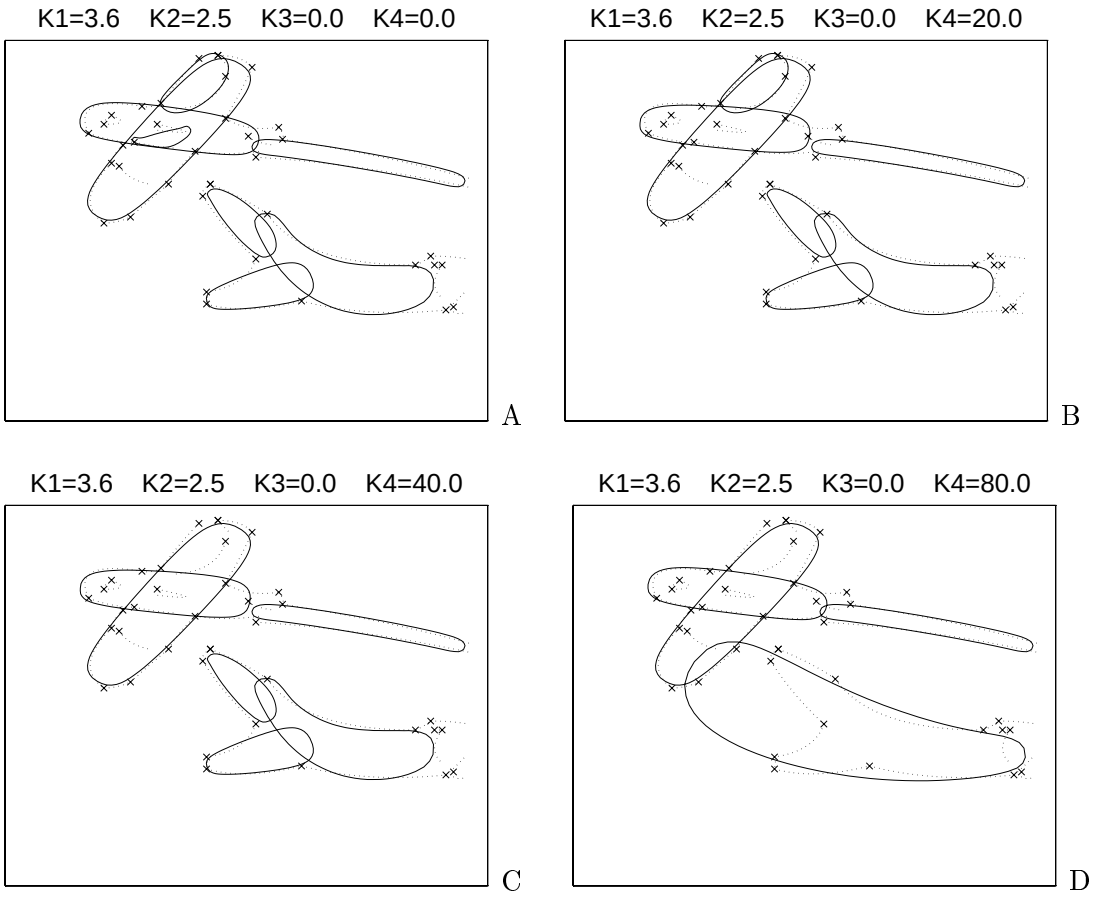


Figure 5.23: Filtering experiments by MDL for the screw-driver, stick and marker example with different values of the model overhead K_4. When K_4 is small, two (fig. A) or one (fig. B) little PDMs inside the marker appear because somehow they describe portions of the images without much conflict with other hypotheses. Both the outline of the marker and the screw-driver handle and shaft are stably recovered throughout the large range of K_4; this is due to the relatively high perceptual salience of the models, which neither have too many competitors. In the case of the wooden stick, the correct segmentation is achieved until a large value of K_4, when an incorrect hypothesis describing the outer contour of the object is elected as most economical (one model versus three); it must be noticed that the elongation of this latter PDM is due to a poor initialisation and to the fact that it has been attracted by the lower part of the marker.

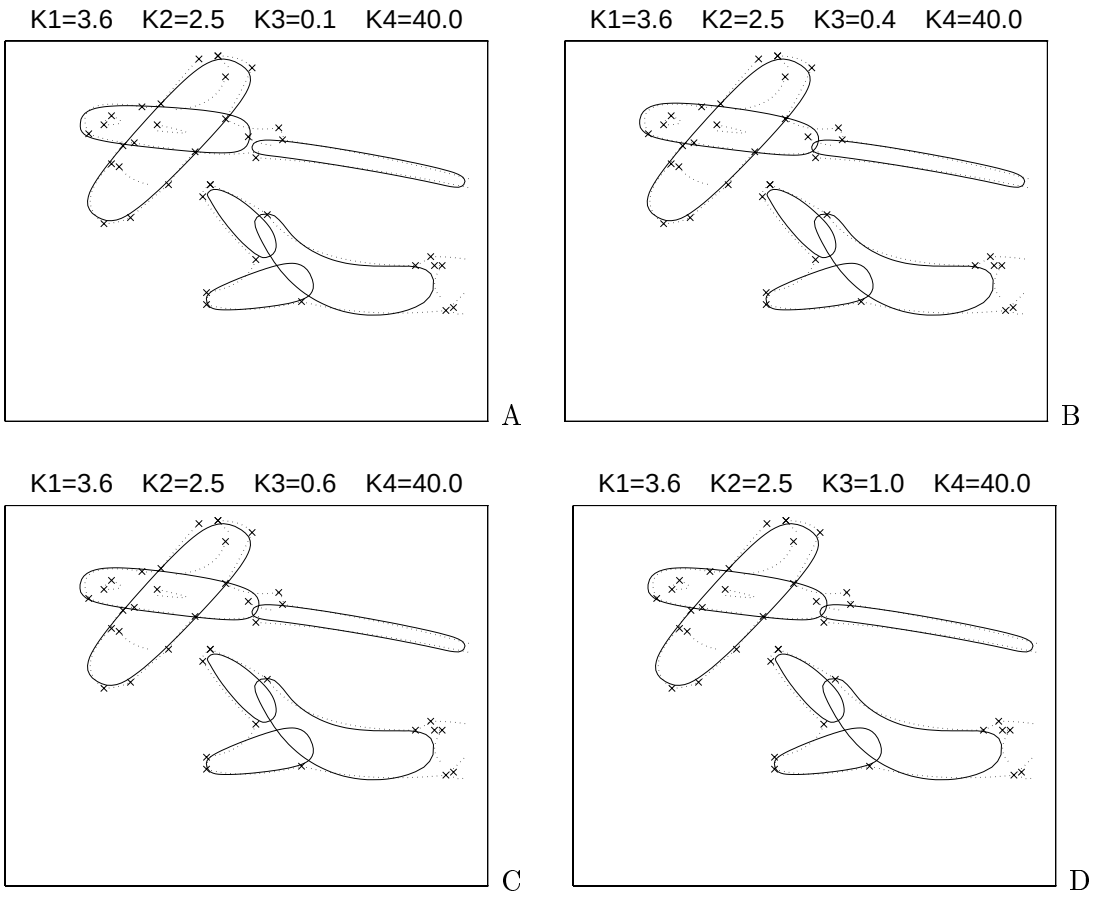


Figure 5.24: Filtering experiments by MDL for the screw-driver, stick and marker example with different values of the residual cost factor K_3, keeping $K_1 = 3.6$, $K_2 = 2.5$ and $K_4 = 40$ fixed. In this case, the stability with respect to K_3 is noteworthy; this can be attributed to the low fitting residuals between model and data that have a small contribution on the overall cost. This situation is very much close to the ones dealt with in [Leonardis *et al.* 94] or [Darrell & Pentland 95], where the residuals were assumed small and Gaussian and hence the high stability of these results is hardly surprising. Note that the two models representing the screw-driver shaft in figs. A and B are slightly different.

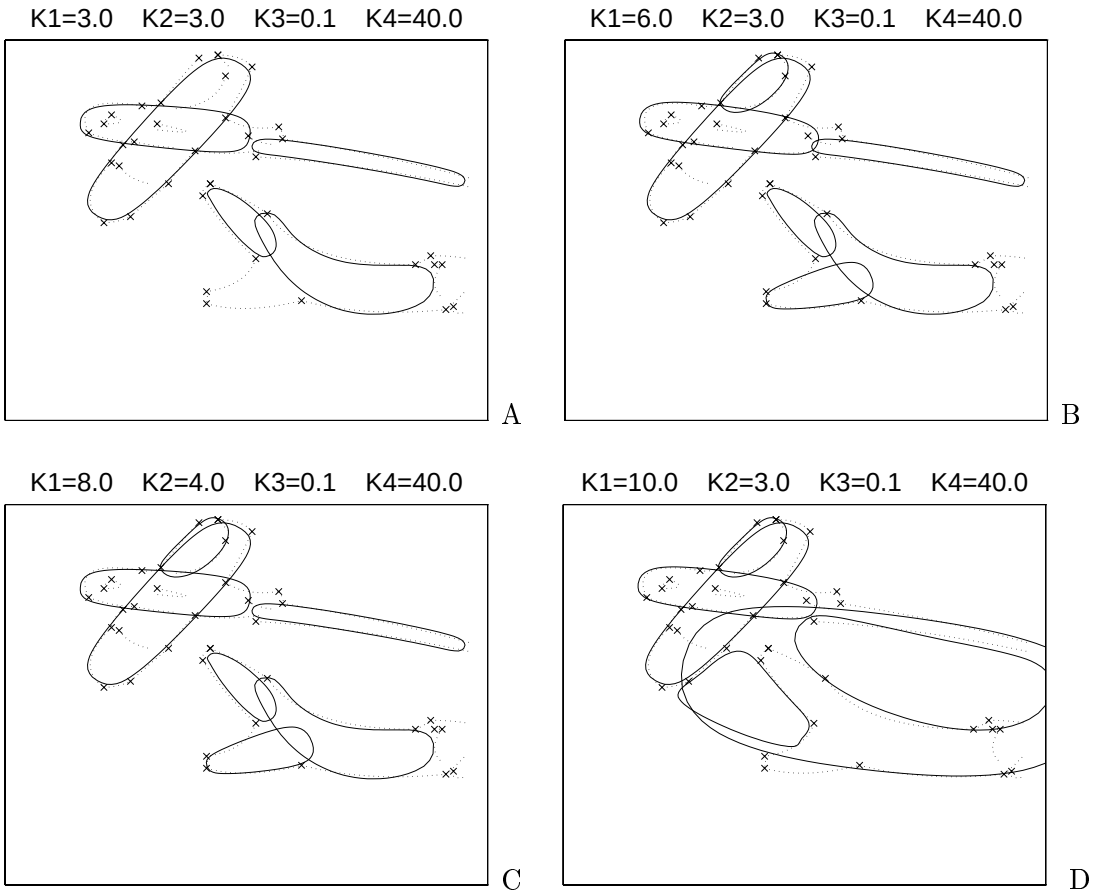


Figure 5.25: Filtering experiments by MDL for the screw-driver, stick and marker example with different values of the constants K_1 and K_2, keeping $K_3 = 0.1$ and $K_4 = 40$ fixed. As seen in Figures 5.23 and 5.24, this example shows high stability with respect to variations of all parameters. However, when the gap between K_1 and K_2 grows too big (fig. D), weird things happen and bigger models tend to appear, as analogously seen in Fig. 5.22-D. In fact, a big K_4 signifies that much weight is given to models supported by as many pixels as possible rather than to ones having a good ratio between supported and unsupported contour portions. The very opposite happens in fig. A, where the lower branch of the wooden stick was not selected because it has too much unsupported contour, as shown by its low salience in Figure 5.7.

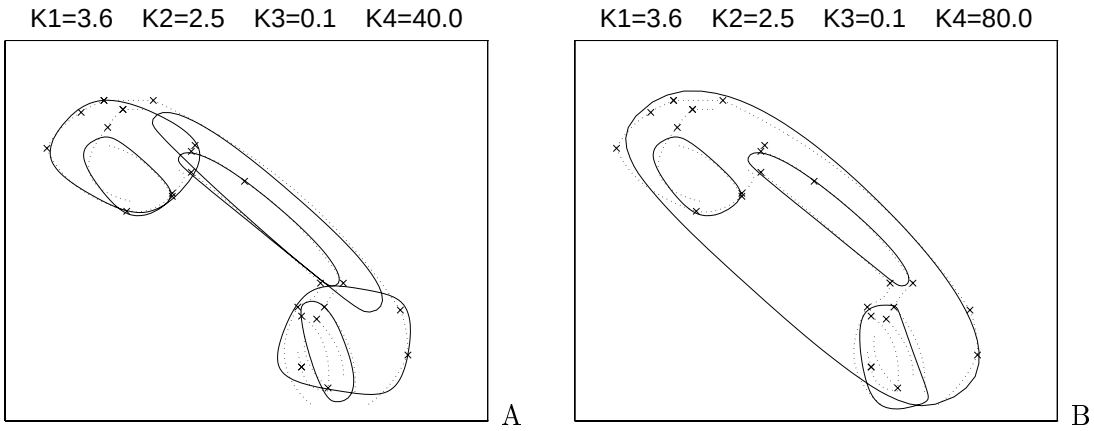


Figure 5.26: Two filtering experiments by MDL for the handset example with different values of the model cost K_4. The handset in this example has a pronounced three-dimensional structure and therefore alternate groupings corresponding to faces are to be expected; this problem is discussed in detail in Section 5.3.7. When K_4 is smaller than about 70, the results are all like the one shown in fig.A. It can be seen that the three main parts (mouth, ear pieces and handle) are correctly selected plus three others corresponding to faces. Because of high cluttering and low resolution, the model selected in the lower piece is rather poor, but yet clearly distinguishable. In the upper piece the selected model actually fits the shadow (see Fig. 4.5) rather than the real part contour; unfortunately this situation cannot be easily avoided by looking just at the edge image. In the case of the handle the smaller elongated PDM fits well the lower face of the prism; the fitting to the top part is disfavoured because of the higher difference between supported and unsupported contours. When K_4 grows big, it becomes expensive to select many models and therefore the big one in figure B that coarsely corresponds to the convex hull of the object is selected; it has to be observed that this is however a very valid representation of the image, since this big model very well matches all the outer contours as well as the three other hypotheses do for the inner edges. The same phenomena for high K_4 was also noticed in other experiments. Experiments for different K_3 have given analogous results as in Figure 5.24, that is the solution remains the same as the one shown here in fig.A.

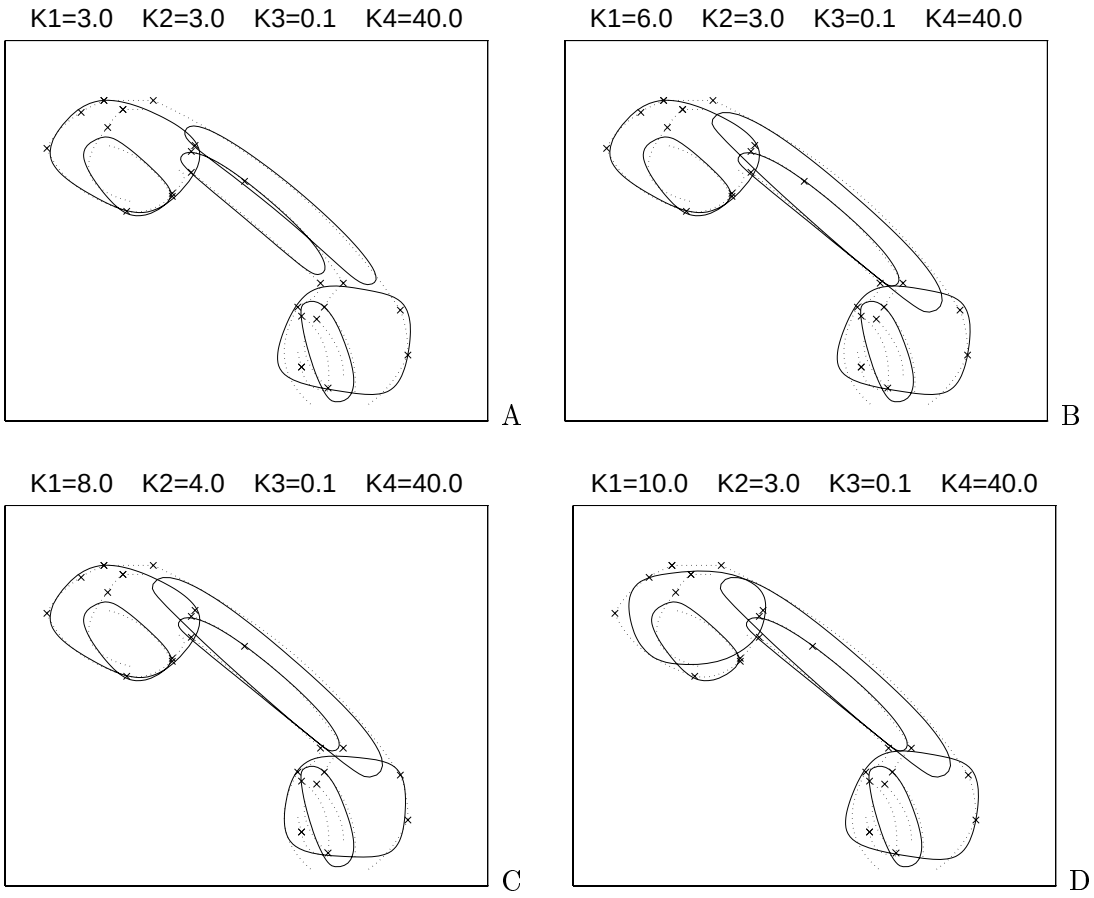


Figure 5.27: Filtering experiments by MDL for the handset example with different values of the constants K_1 and K_2, keeping $K_3 = 0.1$ and $K_4 = 40$ fixed. The results are quite stable in figs.B, C and D. The hypothesis selected in fig.D for the upper piece is slightly different and worse: that could well be a local minimum of the cost function. In fig. A, instead, the two face hypotheses of the handset handle prism are selected; this can be explained by considering that since K_1 and K_2 are equal, high weight is also given to missing PDM contour portions and that solution can be seen as minimising unsupported contours, since the whole-handle hypothesis has more of it. However, this situation is inherently ambiguous and the reasons for this are detailed in Section 5.3.7.

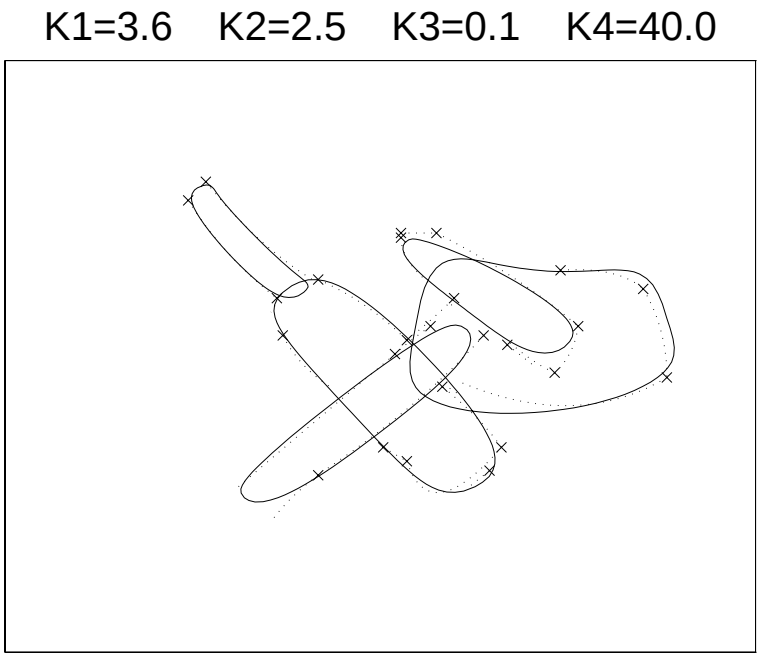


Figure 5.28: Filtering experiments by MDL for the beer bottle and hammer example. This experiment has shown the same kind of stability as the screw-driver, marker and stick example. For very large values of K_4 (> 100) the large object underneath the hammer head disappears and no variations were noticed when changing K_3 as done in the experiment of Fig. 5.24. In addition, when playing with K_1 and K_2 as done in Fig. 5.25, no changes were produced, due to the lack of good competing or ambiguous situations like the one found in the handset example. The figure shows the result for the same values of the constants as the ones that produced good solutions in previous examples, in order highlight that for these four experiments the intuitively correct solutions were obtained with the same parameter configuration.

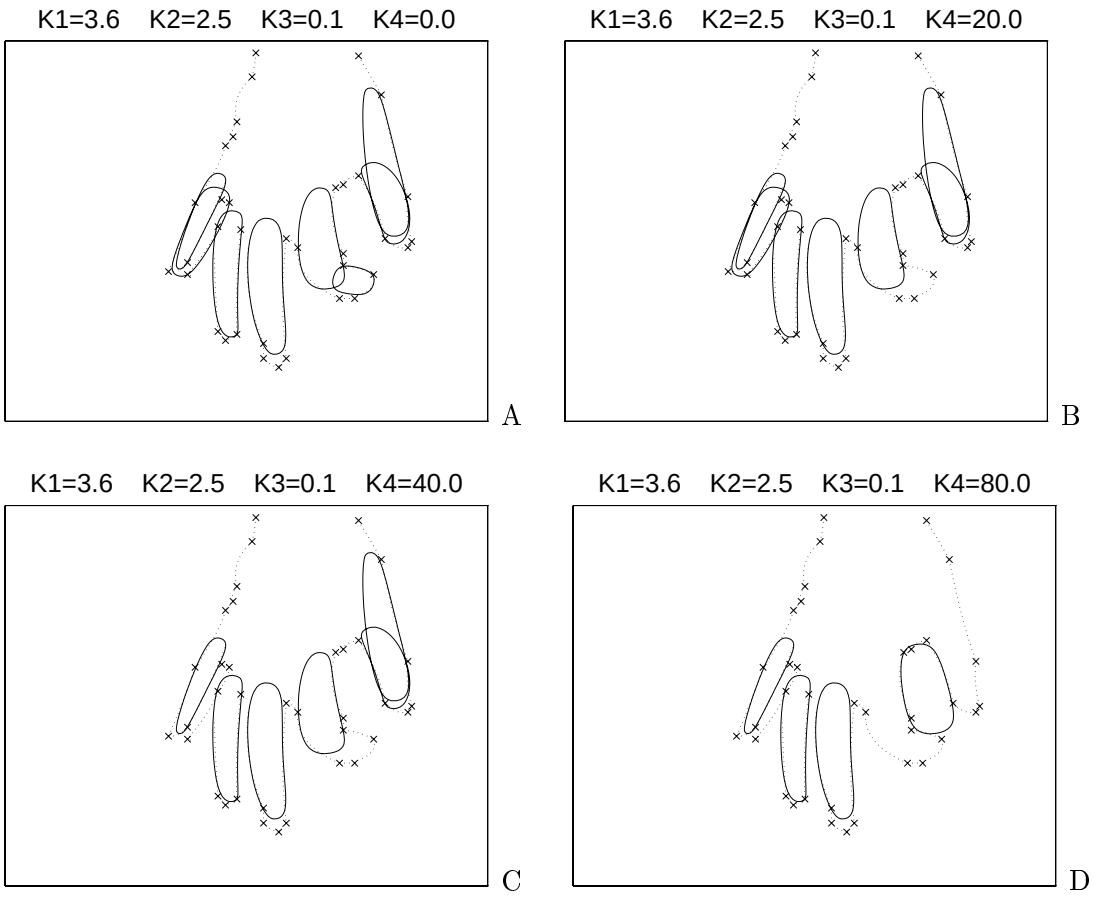


Figure 5.29: Filtering experiments by MDL for the hand example with different values of the model cost K_4. In figures A, B and C good results are obtained. In A and B, due to the low model cost K_4 and a good amount of non-shared support, both the little finger and thumb have double hypotheses; the double thumb is found up to $K_4 = 40$ whereas the last segment of the index is lost soon, since it describes very little contour of the image. The most interesting phenomenon is illustrated in figure D for $K_4 = 80$: index and thumb disappear and leave a background hypothesis that has very high salience. Section 5.4 will show that if information about the background is available, this situation would not arise and the right parts would be correctly recovered.

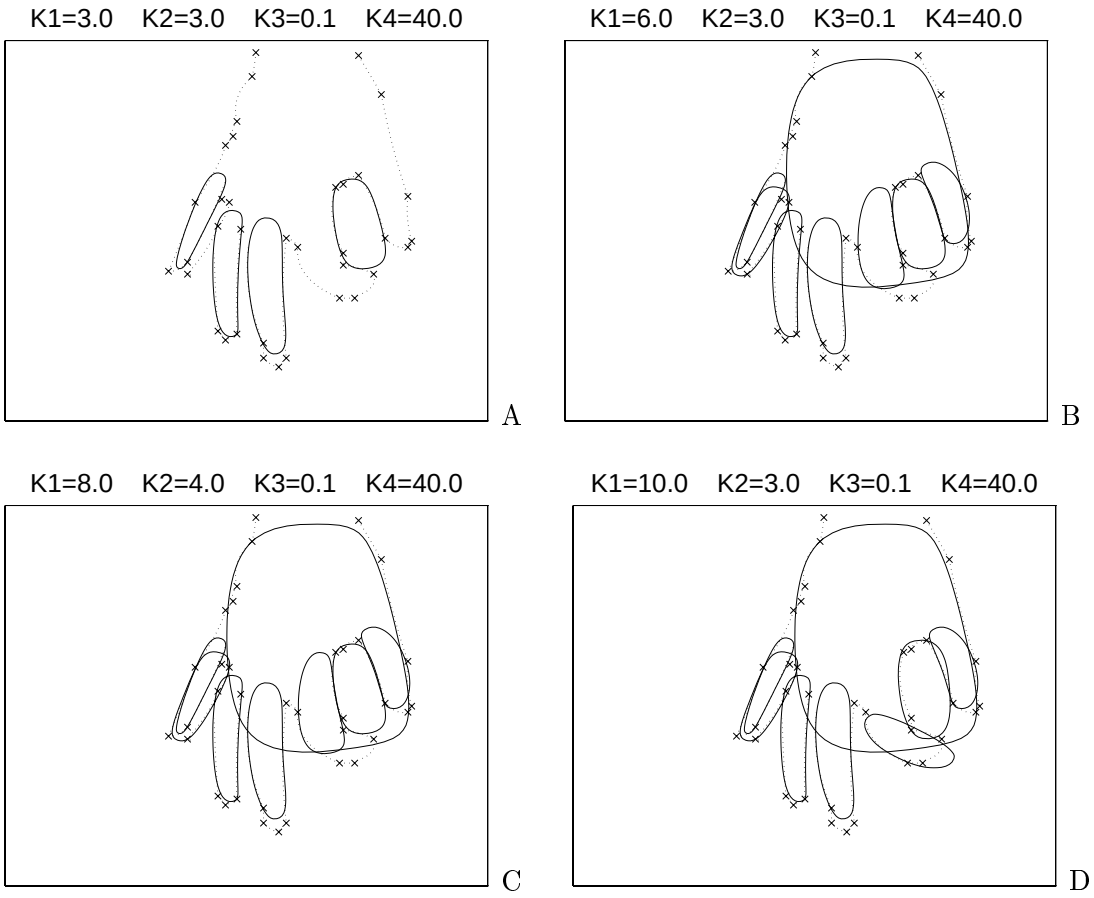


Figure 5.30: Filtering experiments by MDL for the hand with different values of the constants K_1 and K_2, keeping $K_3 = 0.1$ and $K_4 = 40$ fixed to the same values used in previous experiments. When K_1 is much greater than K_2 more models tend to crop up that describe as much contour as possible, with less weight given to unsupported model portions. This behaviour is particularly apparent in figures B, C and D, where both the back of the hand, index, thumb and gap hypotheses are produced. It can be seen that the gap hypothesis does not actually describe much additional contour. In the solutions B, C and D, however, no correct part is missing, apart from the final segment of the index finger. Case A is similar to the one shown in Figure 5.29-D, for which are valid the same considerations made there.

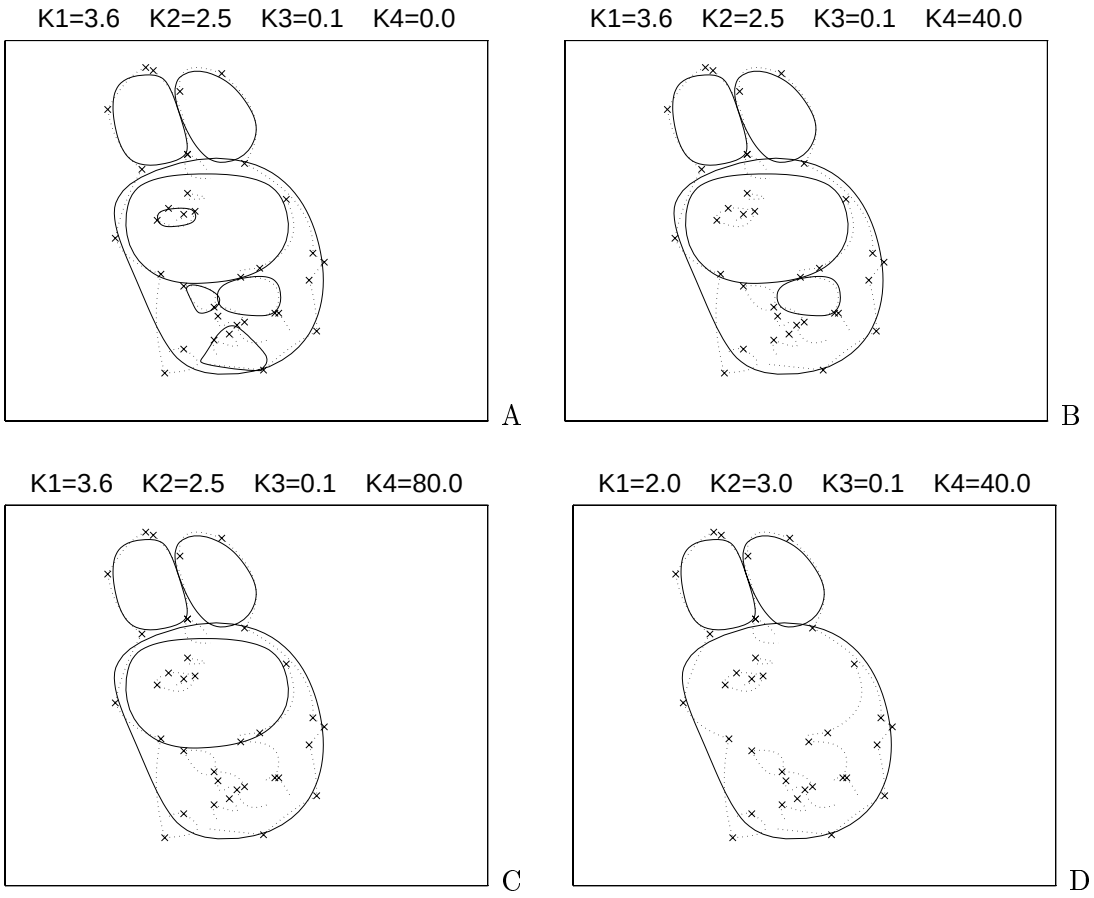


Figure 5.31: Filtering experiments for the toy rabbit example. This example highlights that the method is stable but the determination of the supporting and shared codons is an important factor to be considered. The big hypothesis covers most of the rabbit head and body outline and, due to an unfortunate choice of the threshold parameter, the slightly curved segment crossing the big PDM (at about 1/4 of its left side) was included in its support region. This has caused the body hypotheses (see Fig. 5.13) to never be selected. It must be said, however, that the head-body separation is very subtle, especially due to the shadow extending along the right side of the figure. Apart from this, both ears and head are stably recovered in the experiments in which K_4 was made to vary, that is in figs. A, B and C. The nose and other small details disappear as K_4 grows, as happened in all the other experiments. In fig. D, the head too disappears, due to the unusual choice of the parameters (K_2 should never be bigger than K_1); the head hypothesis, though, was always selected for a rather large range of K_1 and K_2 values.

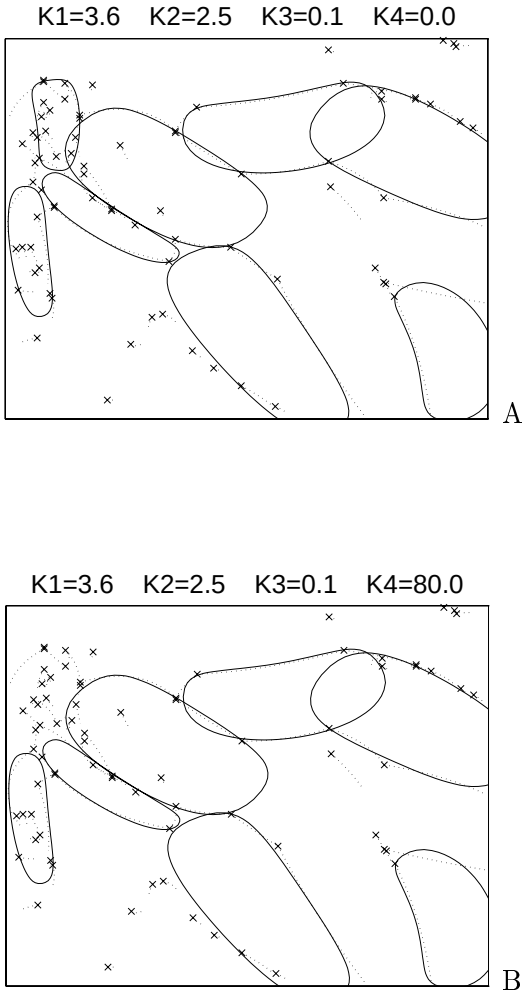


Figure 5.32: A couple of experiments for the complicated Modigliani painting example. This is a very interesting case that almost constitutes a hymn to the impossibility of achieving good part segmentation from edge data only. Although the two forearms, the body and something resembling the head are stably recovered, the two legs could not possibly be selected because of the highly competing hypotheses in the background that not only support the edges of the legs but also the myriad of background edges, in particular the ones at the top. These results do not change by playing with the constants, a fact that indicates that the figure-ground ambiguity here is so strong that the other alternatives are very far below in term of simplicity.

5.3.7 Where do problems come from?

As we have seen in the previous experimental section, the MDL support competition method is an interesting one for our part segmentation task but some problems have surfaced from experimental evidence. Most of these problems can be attributed to the well known figure-ground ambiguity in edge images but some issues, more specific to the MDL method, have been identified and the most significant ones are reported in this section. I believe that this section bears a certain significance because from the recent literature it appeared that the MDL hypotheses support method would be able to handle in principle several middle-level segmentation problems.

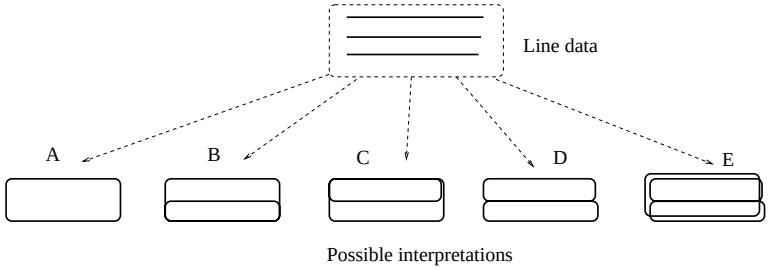


Figure 5.33: Taxonomy of possible MDL filtering results for the case of three parallel lines. (A) Only the bigger hypothesis is selected, which normally corresponds to the actual part outline. (B, C) The bigger one plus either of the small ones are selected, as happened in the case of the handset in Figures 5.26-B, C and D. (D) Both small hypotheses are selected, as happened in Figure 5.26-A. (E) All three hypotheses are selected. The most common ambiguities arise for cases B, C and D, as described in the text.

Figure 5.33 shows how within the proposed framework, three parallel lines can give rise to an ambiguity in terms of their interpretation by models. This situation arises often in images with multiple objects and it is even accentuated when parts of the line are missing because of occlusion or poor edge detection.

Let us suppose for a moment that the error of fit is zero and let a be the length of each line and b the distance separating them. Then, for the five cases in Figure 5.33, the

overall bit saving S as expressed by Eqn. (5.6) through Eqn. (5.7) and (5.7) becomes:

- Case A $S_A = 2K_1a - 4K_2b - K_4$
- Case B $S_B = 3K_1a - 6K_2b - 2K_4$
- Case C $S_C = 3K_1a - 6K_2b - 2K_4$
- Case D $S_D = 3K_1a - 4K_2b - 2K_4$
- Case E $S_E = 3K_1a - 8K_2b - 3K_4$

It should be clear that by changing the constants K_1, K_2 and K_4, the highest values of S can be obtained in diverse situations. Case D is slightly favoured over B and C (less non-supported contour portions) but unbalanced lengths and some cluttering could favour either (see the handle of the handset example). Case A could be favoured if a is small and therefore the saving due to the description of the edge by the models cannot compensate for the model overhead itself (e.g., see the final segment of the index in the hand test image). Case E should never happen, and I have not seen it in the experiments I have conducted.

Similar problems arise in cases such as Figure 5.34, which is inspired by two parallel fingers with incomplete edge data. The desirable solution is the one indicated by A, since since both models in solutions B and C have a higher percentage of unsupported contour, but the presence of occlusion, missing data, and badly fitted models could favour either of the two other solutions. One might argue that cases A, B and C could be easily disambiguated by properly penalising overlapping; this is certainly true and some experiments have been carried out in this regard. However, by doing so the possibility of dealing with occlusions – that is one of the strong points of the approach to part-based grouping presented in this work – would be precluded. This problem could only be solved by using more contextual information or by the integration of other information, as we shall see in Section 5.4.

Although the MDL method is in principle very stable, situations like the above can lead to some instabilities, as we have seen in the examples in the previous section. Ambiguities and instability of this kind have often not previously been noticed in related

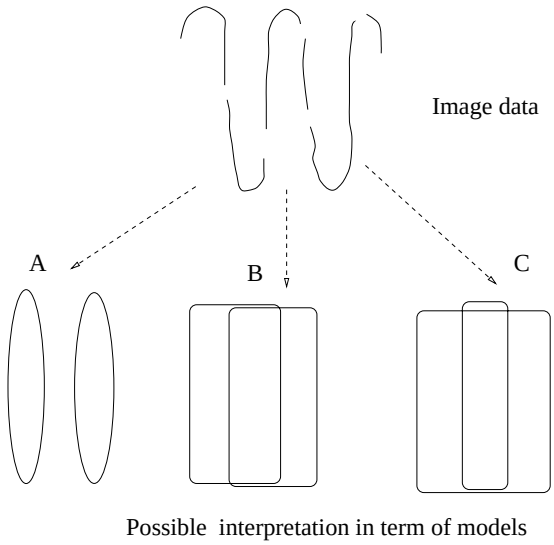


Figure 5.34: Another situation that might lead to instability of the results. The case is inspired by two parallel fingers of a hand. At the top, the example image data is shown that has some missing boundary portions. The correct solution is exemplified in figure A. In figures B and C two equally good solutions, in term of accuracy of contour description, are shown. This kind of ambiguity can always arise because the hypotheses in B and C are always produced as well as the correct ones (see, e.g. the set of hypotheses in Figure 5.4) and can only be avoided by employing additional information or by high level knowledge.

literature[8]. Tuning the parameters on a simple test example to favour a particular solution would clearly be useless because in real conditions even a small percentage of missing contour would turn the balance towards alternative solutions.

Another well known problem of the MDL framework regards the choice of parameters – the perennial problem of computer vision – and this limitation was pointed out as early as in [Witkin & Tenenbaum 85]. However, in favour of the MDL paradigm and against other heuristic methods it must be said that, although these constants are hard to determine, good ranges are experimentally found that are always within ranges that can be predicted with a certain confidence, as shown in Section 5.3.4. For instance good values for K_3 are always between 20 and 60, which can well be the number of bits necessary to express the model.

Finally, there is the problem of scale. Although it has been advocated elsewhere

[8] Interestingly enough, similar problems have been reported in robotics literature in, e.g., [Miglino *et al.* 96].

[Pednault 89, Leonardis *et al.* 95] that one of the main advantages of MDL is its scale-independence, practically this is not true because the number of bits necessary to describe supported and unsupported regions and residuals are absolute numbers that depend on the model dimension; in fact, it has been observed that bigger model hypotheses were slightly favoured over small ones (such as the last segment of the index finger in the hand example) because they lead to higher savings in bits. These scale problems have actually been considered assets in works such as [Leonardis *et al.* 95] and [Darrell & Pentland 95] because bigger models were supposed to more economically describe the surface being segmented under their Gaussian noise axioms. A simplistic solution could be to tie K_4 to the model scale but this not only would contradict one of the main assumptions of the MDL method but has no theoretical support; further work has to be done in this regard.

5.4 Integration of more information: Knowing the background

As we have seen in the experimental section, in particular in the hand and Modigliani painting examples, high-scoring part hypotheses corresponding to the background can lead to a wrong part-segmentation. In this section we provide an example that illustrates how additional information on the background can be used in order to let the MDL filtering method produce the right hypotheses where strong figure-ground ambiguities are present.

The example used here is the hand test image. Only hypotheses that have at least about 40% of their area belonging to the foreground are considered for the MDL filtering stage (Figure 5.35-A); the remaining background hypotheses are rejected (Figure 5.35-B). Note that there is no need for this stage to be precise: the only important factor is that parts that are *very likely* to belong to the background are rejected.

Here, the selection has been performed by hand but this could be easily done automatically, once the information on the background is available through e.g. thresholding of colour, depth information or texture characteristics.

Figure 5.36 shows a comparison of the results. With the parameter configuration indicated in the figure, when all hypotheses are used indiscriminately the gap between thumb and index finger takes over the correct part hypotheses (Fig.5.36-A). When the background hypotheses are not considered, the correct solution is produced, as shown in Figure 5.36-B.

Clearly, the rejection of background hypotheses trivially applies also to the filtering by saliency thresholding presented in Section 5.2, in the sense that the hypotheses belonging to the background would simply not be included in the final solution.

Additional information (not only regarding the background) could be more formally integrated through the explicit use of Bayes probabilities in a framework such as, e.g., [Sakar & Boyer 93], but this is out of the scope of this thesis and is left as future work.

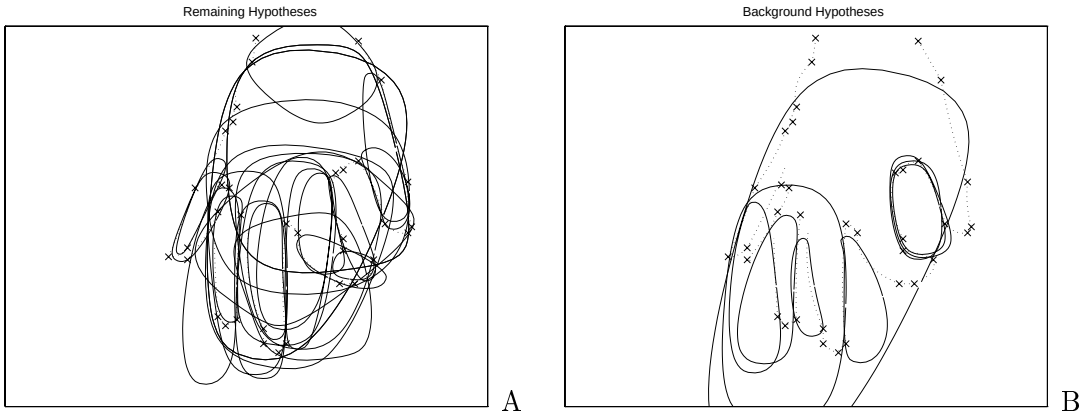


Figure 5.35: Background and foreground hypotheses for the hand example. Hypotheses that have at least about 40% of their area belonging to the foreground have been selected for the MDL filtering stage(figure A); the background hypotheses are displayed in figure B. The selection has been performed by hand but it could be easily done automatically once the information on the background is available.

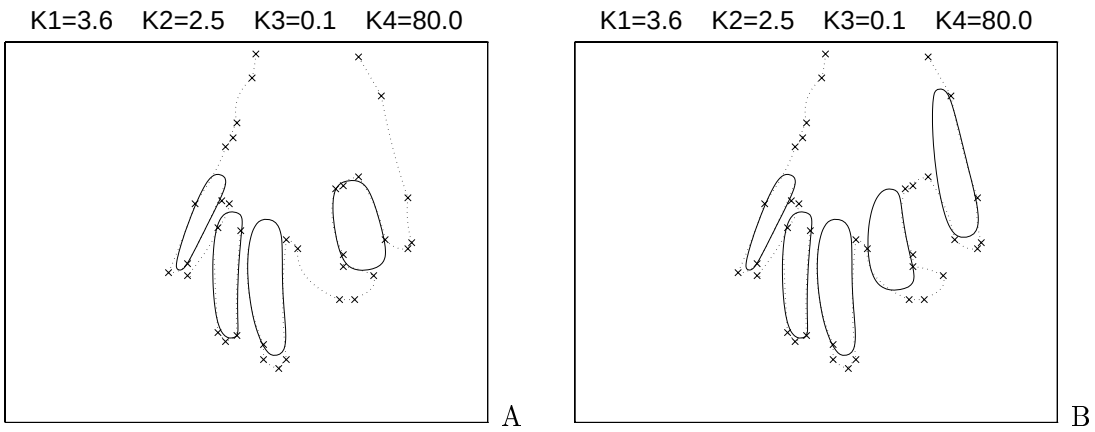


Figure 5.36: Filtering results by excluding background hypotheses. With this particular parameter configuration, when all hypotheses are used indiscriminately, the gap between thumb and index takes over the correct part hypotheses (fig. A). If information on the background is used and hypotheses with very high probability of belonging to the background are not included in the filtering, the correct solution is found, as shown in fig. B.

5.5 Discussion

In this chapter two methods have been described that allow the initial set of hypotheses produced as in Chapter 4 to be filtered down to few ones that have a high likelihood of corresponding to actual parts of objects.

The first filtering method proposed in Sec. 5.2 is hardly new. It employs straightforward thresholding of a very simple saliency measure; the experiments showed that it is very hard to obtain minimal description with this method and the final set of selected hypotheses is often highly redundant.

Rather than a viable proposal, this first method was presented as a prelude of a more sophisticated method that accounts for competition between multiple representations. The novel proposed method, presented in Sec. 5.3, is inspired by recent work in surface segmentation by the Minimum Description Length principle [Leonardis *et al.* 95, Darrell & Pentland 95].

It is now worth pointing out that we have carried out a psychological experiment in order to informally assess whether humans would give similar judgements when presented with the same edge images. Results are – as expected, I would say – quite encouraging. The notion of part as a MDL representation of many objects is a sensible one, both from a strictly engineering [Pentland 86] or cognitive [Rosch 73a, Biederman 87] point of view: the essential structure of objects is better described by assemblies of simple parts rather than by an holistic multiple view-based description. Many researchers denigrate this view, seeing it a too simplistic [Tarr & Bulthoff 95]. However, in my opinion, the real question is whether our description of things by words has a correspondence in the phenomenology of the visual processes, but philosophising on this would take us far beyond the scope of this thesis. The full description of the experiment along with results and further considerations are given in Appendix E.

In the following two subsections, contributions of this new method are briefly recapitulated and future extensions proposed.

5.5.1 Contributions

The relevant contributions of this chapter are the following:

- A segmentation strategy proposed in [Leonardis *et al.* 95] and [Darrell & Pentland 95] has been extended from the three-dimensional to the two-dimensional domain. The method was shown to produce good results in 3D surface segmentation and an application to part segmentation was also proposed in [Leonardis *et al.* 94]. Here the method is for the *first time* used to perform part segmentation from ordinary unsegmented edge images and many experiments are described that show the validity of the approach.

- In the 3D case, the hypotheses are produced by simply fitting and growing quadrics to range data starting from randomly placed seeds; however, the MDL hypothesis competition method was never applied to the part (or feature) segmentation problem of 2D images because it was not clear how the large number of hypotheses including the right ones could be generated. The hypothesis generation method described in Chapter 4 has managed to do so and when used with the present filtering method it can be considered as one of the very few, if not the only, method for fitting deformable models (in our case the generic part PDM) to unsegmented real 2D images. To the best knowledge of the author, only the excellent work of [Dickinson *et al.* 92b][9] has done so, albeit with a completely different strategy.

- Some theoretical limitations of the MDL method, e.g. the problem of ambiguities described in Sec. 5.3.7 which have not pointed out in previous work, have been discovered and discussed. These problems are inherent in the use of sole edge information but could also arise whenever the data is incomplete, cluttered or the fitting residuals are simply too high and no noise model is available.

- A lesser contribution regards the use of a genetic algorithm to maximise the quadratic boolean cost function in Eqn. (5.6). In other works, this maximisation was performed by greedy methods but, through some experiments performed by

[9] This method is reviewed in Chapter 2.

simulated annealing, it has been found out that those semi-local approaches could not perform in cases where a high number of strongly competing coarse hypotheses have to be sorted out. In [Darrell & Pentland 95] and [Leonardis *et al.* 94] it was possible because their competing hypotheses were most of the time both correct and with low fitting residuals, a situation that occurs seldom in cases such as that dealt with in this work.

5.5.2 Future extensions

First above all, the support finding method presented in Section 4.6.3 is of a rather heuristic nature and perhaps more work could be done in making it more scale independent and for including some heuristic that would account for what we perceive as support.

As a matter of fact, we have so far used codons as atomic entities and referred to them as the sole source of information but this assumption, although yielding good results, is rather simplistic. Codons could in fact be dropped at this stage as image support is directly found in the raw edge or gradient image, which would probably avoid some of the problems that have been encountered in choosing thresholds for the determination of codon support.

Another significant improvement concerns the integration of more information (such as region, colour and depth) under the MDL framework. Thus far, no work has been done on merging different kinds of information under an MDL framework but this could be just incidental, because a clear derivation could be made from a Bayesian frame of mind (e.g. [Sakar & Boyer 93]) in which each piece of information is subject to a conditional probability.[10] This procedure may remove some ambiguities, such as those reported in Section 5.3.7.

Finally, a few words should be given about an intriguing possibility. The presence of ambiguities that give similar values of the cost function can also be considered as mutually exclusive interpretation of the same image. By employing a simple genetic

[10] It should be said that this integration is somewhat in contrast to the philosophy of the MDL principle, where one finds the best interpretation in terms of a formal descriptive language; arguably, however, the descriptive language could be made more complicated in order to encompass multi-valued data.

algorithm such as the one described in Section 5.3.5, only one of the many alternative solutions survives the evolution and others tend to extinguish. If a *multi-population* GA were used instead (e.g., [Fuger *et al.* 94]), all these equally high scoring alternative solutions could be allowed to evolve (up to migrations) in parallel. A similar possibility was also mentioned in [Hill & Taylor 92].

Chapter 6

Fitting Parametrically Deformable Aspects

This chapter[1] proposes an approach to the fitting of generic solid parts to unsegmented edge images.

The structure of the chapter is as follows. First, an introduction to the approach and a brief review of previous related research is given. Next, the *parametrically deformable aspect* construction is described, followed by Section 6.5 on the fitting procedure. The experimental system and the aspect-based control strategy are outlined in Section 6.6. Section 6.7 presents and discusses some experimental results that show the validity of the approach but also its limitations. The chapter concludes with a discussion on the contributions, restrictions and some future work is suggested.

6.1 Overview

As introduced in Section 2.3, geons [Biederman 87] are generic solid primitives defined by qualitative properties of the axis and cross-section of a generalised cylinder[Binford 71] that are invariant under change of viewpoint.

In the previous chapters, it has been shown that qualitative 3D primitives like geons can be segmented out from real images by looking for their outline but the essence of their 3D structure (the *geon class*, according to [Biederman 87]) is lost in the process.

[1] An earlier version of this chapter that did not use the aspect-based control strategy appears in [Pilu & Fisher 96d]. The whole chapter is available as a technical report in [Pilu & Fisher 96e].

For instance, in the part segmentation of the handset of Figure 5.26-A, both pieces and handle hypotheses have been correctly produced but the geon class cannot be recovered except by high-level reasoning about the hypotheses' layout.

In this chapter, a new method is presented for fitting qualitative 3D volumetric parts models to real 2D images that treats geons[2] as *single entities* to be extracted from images. This is done by matching *parametrically deformable contour models* (PDCMs) of geons to edge images in the framework of Model-Based Optimisation (MBO), in which an objective function expressing the global likelihood (goodness) of fit is maximised. The cost function accounts for both matched and unmatched contour portions and is formulated in Bayesian terms.
The potential advantages of such a global approach lie in imposing overall consistency on the image which lead to robustness to cluttering and opens possibilities of direct figure-ground segmentation in the spirit of [Leonardis *et al.* 94] or the method presented in the previous chapters.

Similar approaches to generic part recognition that used deformable superquadrics as generic shape models have been investigated for the 3D case (range data input) in popular works such as [Solina & Bajcsy 90] and also in [Wu & Levine 94], [Leonardis *et al.* 94] and [Borges 96]; only in [Metaxas *et al.* 93] was the method extended to the 2D case as a front-end of the OPTICA system [Dickinson *et al.* 92b]; other related methods are reviewed in Section 2.4. To date, however, one of the main problems faced by global fitting approaches is their sensitivity to the initial state of the models, which often compromises the quality of the solution. In previous work [Pilu & Fisher 96d], we used a loosely-constrained optimisation approach which worked well only when the initial model was topologically equivalent to the geon instance being fitted. Here, this deficiency is reduced by using an aspect-based hypothesis generation-and-testing strategy inspired by [Eggert *et al.* 95]. The multidimensional parameter space defining the geon PDCM is partitioned into eight topology-equivalent classes which have been called *parametrically deformable aspects* (PDAs); the set of eight PDA can be seen as a single deformable model endowed with global topology information. By doing so, the optimisation can independently focus on regions of the parameter

[2] The parts will be still called geons, although they are a subset of the ones defined in [Biederman 87].

space that correspond to models with the same topology, thereby reducing the chances of getting stuck in local minima caused by different interpretations of image features. A simple experimental control strategy suggested by [Eggert *et al.* 95] is employed that, by starting from coarse 2D part hypotheses produced as in the previous chapter:

(1) initialises all eight PDAs at a representative position for each PDA;

(2) performs the fitting independently for each PDA thus initialised;

(3) chooses the one that achieves the best score.

We will see that the happy marriage between parametric deformable contour models and the concept of topologically different aspects efficiently represents geons and yields more robustness in the optimisation process we use, which is Simulated Annealing.

The results we achieved from 2D images are very much comparable with the one obtained by using 3D range data (e.g. by [Solina & Bajcsy 90]), although depth and orientation obviously cannot be recovered from 2D images.

6.2 Review of previous related work

In this section, some previous research in model-based optimisation and the use of aspects in recognition are reviewed.

6.2.1 Model-Based Optimisation

In the context of computer vision, Model-Based Optimisation (MBO) aims at finding the best fit of a model by minimising an objective function (or maximising a likelihood function) that can incorporate both high and low level knowledge about the image, object model and goodness of fit. Within this framework, the use of whole boundary models – such as the one used here – is the most natural and effective because [Staib & Duncan 92]: *i)* the whole structure is imposed on the problem and the task is simplified; *ii)* gaps are naturally filled and *iii)* overall consistency is more likely to result.

MBO can be performed in parameter space or in image space and with arbitrary

models, fixed templates or deformable models.

Optimisation in image space is done through fitting each composing element (point, lines, etc) of the model more or less separately to the image. Typical models that have been used within this paradigm are fixed templates [Eggert *et al.* 95], arbitrary models like snakes [Kass *et al.* 88], bead chains [Critton & Parish 83], Markov boundaries [Friedland & Rosenfeld 92] and parametric shapes like Point Distribution Models (PDM) [Cootes *et al.* 94]. As we shall see later, this method allows the model to better track object irregularities but, besides problems of stability, it is often difficult to incorporate high-level knowledge about the overall object shape to guide the fitting process. In most works using these types of models, the high-level knowledge is inspired by physical analogies (such as the smoothness constraint [Kass *et al.* 88]) but very promising results have been achieved by using PDM [Cootes *et al.* 94] or finite element models [Pentland & Sclaroff 91], where global information is encoded in the modes of variation.

On the other hand, MBO in parameter space is performed by adaptively changing the parameters of the model and checking the goodness of fit in the image; it implies the use of parametric (deformable) models whose shape variability can be expressed in a compact form by few significant parameters; within this paradigm, there are works such as by [Lowe 91], [Yuille *et al.* 92], [Staib & Duncan 92] and a wealth of others. Fixed templates have also been used in this context but that is a sort of degenerate (though important) case in which the only controlling parameters are those defining the pose of the object. Although the use of parametric models offers the advantage of compactness of representation and easy classification, often the optimisation in parameter space turns out to be a hard problem (see, e.g., [Lowe 91]), also because the parameter space is often not as "tight" as for arbitrary models.

As far as the optimisation algorithm goes, that is the "tactic" for finding the best fit in terms of configuration or parameter values, several methods have been proposed and experimented with but none has provided a reliable and sufficiently general method. Cost functions are often strongly non-linear and present many possibly narrow and shallow local minima that make fast convergence hard even to a sub-optimal minimum. The initial condition, that is the values of the model parameters before the optimisation

starts, often plays a crucial role and, where this is not done manually, several heuristics have been timidly proposed (such as in [Lowe 91] or [Friedland & Rosenfeld 92]). Commonly used methods include Levenberg-Marquandt (used, for instance, in [Lowe 91] and [Borges 96]), Simulated Annealing [Wu & Levine 94] and hill-climbing in combination with continuation or multi-scale techniques [Staib & Duncan 92].

6.2.2 Use of aspects

The concept of *aspects* was first formulated in [Koenderink & vanDoorn 79] and a new object representation, called the *aspect graph*, was proposed. An aspect graph is essentially a *"...complete enumeration of topologically distinct views of an object, along with a definition of the region (*cell*) of viewpoint space from which such a view is seen"* [Eggert *et al.* 95].

A number of algorithms have been proposed to compute the aspect graphs of polyhedra [Stewman & Bowyer 90], algebraic surfaces [Ponce *et al.* 92], suggestive models [Fitzgibbon & Fisher 92] or solids of revolution [Eggert & Bowyer 93], often by approximating the exact solution by tessellating the Gaussian view-sphere. However, the practical use of aspect graphs for recognition has been hindered by the lack of practical implementations and therefore they have been mainly used for feature prediction, that in for checking how a feature combines with others. Relevant works that used such an aspect-graph based recognition strategy are, for instance, [Dickinson *et al.* 92b], [Chen & Kak 89] and [Ikeuchi 87].

A major conceptual extension of the use of aspect graphs has been proposed in [Eggert *et al.* 95] where the distinct-topology property of aspects is used to constrain an iterative fitting method within a single view-cell, thereby dramatically improving convergence quality and speed.

In most aspect-based works, including [Eggert *et al.* 95], CAD models were used because of the difficulty of constructing aspect graphs for general smoothed objects. One of the major contributions of this chapter is to show that the use of an aspect based strategy is very beneficial also for the fitting of generic deformable models, such as superquadrics, in which a "topology-blind" strategy often yields poor results.

6.3 Parametrically deformable contour models of geons

Geons are volumetric shapes that are defined by qualitative features and are hence subject to high intra-class variability. Within our framework of Model-Based Optimisation, the recognition of geons from 2D images needs to have a model that can describe in a compact way their projected contour and, because geon models computed inside the innermost loop of the optimisation process, this must be done as speedily as possible.

Recent works that dealt with the recognition of geons from 3D range data (e.g. [Solina & Bajcsy 90], [Borges 96], [Wu & Levine 94], [Raja & Jain 92b]) have associated geons to globally deformable superquadric (DSQ) model [Barr 81]. There are mainly two advantages in using DSQ models. Firstly, the distinguishing features characterising geons can be expressed by single parameters such as bending, roundness, swelling and tapering and, secondly, they can represent very compactly a variety of shapes [Pentland 86].

In this chapter, the use of superquadrics in extended, in the spirit of [Metaxas *et al.* 93], to the 2D case by approximating the contour of the image projection of geons (as opposed to their spatial occupancy) by the apparent contour (outline plus interior edges) of globally deformable superquadrics once they have been properly deformed, roto-translated and projected onto the image.

Unfortunately, computing the apparent contour of DSQ and in general for smooth surfaces is not a trivial job. As classic works in aspect computation show [Eggert & Bowyer 93, Petitjean *et al.* 92], if an exact closed-form solution is sought, huge systems of equations need to be solved and time-expensive search in high-dimensional hyper-spaces has to be carried out. For these reasons, I did not proceed along this avenue, which has been followed by the (nevertheless excellent) work by [Metaxas *et al.* 93], where the superquadric contour was found by numerical methods at a high computational expense.

A few words must be said about this use of DSQ. Although they are a good model for representing 3D shapes, they are extremely clumsy mathematical toys. Their deformations are a bit of an engineering hack and their error of fit function has no closed form [Solina & Bajcsy 90]. In particular, when used to compute contours as done in

[Metaxas *et al.* 93], the clumsiness of superquadrics is certainly too much a burden for the compactness of representation they can give in exchange.

However, for our purposes, there is no need to have a precise knowledge of the projected DSQ contour for the following reasons:

- Contour details such as cusps cannot be reliably detected in real images and, if recovered, they would be useful only for structural analysis of the contour such as done in [Bergevin & Levine 93], which has been proved inapplicable in real cases;
- Very few, if any, actual geon-like parts can be properly described by DSQ: they are an arbitrary approximation in the first place, and a *different* approximation does no harm;
- If precise means expensive, the above reasons are even stronger.

Therefore, a new, more straightforward approach has been followed that uses a parametrically deformable contour model (henceforth PDCM) that simulates the deformable superquadric contour in a much more efficient way by explicit construction; this constitutes a significant efficiency improvement to the model building method used by [Metaxas *et al.* 93], where the projected contour of a DSQ was computed numerically at a considerable computational expense.

The geon PDCM has been designed following the pragmatic spirit of [Yuille *et al.* 92], [Cootes *et al.* 91], or [Ferryman *et al.* 95], where models are designed with recognition in mind, rather than being *inherited* from computer graphics or the mathematics literature, as in the case of superquadrics. For instance, in [Ferryman *et al.* 95] a parametric 3D wire-frame model of a car was purposely built that was able to represent the essential shape of several vehicle classes through its parameters; the 2D projection was trivially obtained from the 3D model and the fitting was performed using the technique presented in [Cootes & Taylor 92] and also used in this work. The approach is rather pragmatic but, if "theoretical" support is sought, it fits in the philosophy of [Witkin & Tenenbaum 85], which advocated that *vision has to be driven by structure.*

The model, that is going to be described in the following, is suitable for qualitative

geon PDCM and simulates the contour of projected deformable superquadrics in a very efficient way: starting from a cylinder centred on the z axis with superelliptical cross-section (Fig. 6.1-left), we apply deformations and rotations and find the contour by trivial geometric considerations. In the following the construction of the model is detailed.

The initial superelliptical cylinder $\mathbf{S}$ of height $2 \cdot a_z$ and semi-axes a_x and a_y can be expressed as[3]

$$\mathbf{S} = \begin{bmatrix} \mathbf{x}(\eta) \\ \mathbf{y}(\eta) \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} a_x cos(\eta)^\epsilon \\ a_y sin(\eta)^\epsilon \\ \mathbf{z} \end{bmatrix} \qquad \begin{matrix} -\pi \leq \eta \leq \pi \\ -a_z \leq \mathbf{z} \leq a_z \end{matrix} \tag{6.1}$$

where $0 \leq \epsilon \leq 1$ controls the degree of squareness of the cross-section from a rectangle for $\epsilon \to 0$ to an ellipse for $\epsilon \to 1$.

Any curve lying on this cylinder can be variously deformed but for our purpose of representing geons we are particularly interested in three kinds of deformations: *tapering*, *bending* and *swelling* along the principal axis. Below, the mathematical definitions of these deformations are given. The tapering and bending deformations have been derived from [Solina & Bajcsy 90] but the latter has been slightly modified by normalising the bending control parameter to a_z and allowing bending on both sides which has also improved the stability of its estimation. The swelling deformation, however, has been introduced here to represent the "expanding and contracting" sweeping rule (Fig. 2.2).

Let us indicate by $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ the vector of shape points before and after the deformations, respectively.

A *linear* tapering deformation along the z axis is given by

$$Taper(K_x,\ K_y,\ \mathbf{S}) = \begin{cases} \mathbf{X} = (\frac{K_x}{a_z}\mathbf{z} + 1)\mathbf{x} \\ \mathbf{Y} = (\frac{K_y}{a_z}\mathbf{z} + 1)\mathbf{y} \\ \mathbf{Z} = \mathbf{z} \end{cases}$$

where $-1 \leq K_x \leq 1$ and $-1 \leq K_y \leq 1$ express the amount of tapering in the x-z and x-y plane, respectively; henceforth we shall assume $K_y = K_x$.

[3] Although using similar terminology, the deformations defined here are the extension to 3D of the ones presented in Section 3.3.

A *circular* bending deformation in the y-z plane is obtained by (see [Solina & Bajcsy 90] for details):

$$Bend(c,\ \mathbf{S}) = \begin{cases} \mathbf{X} = \mathbf{x} + sign(c)(R' - r) \\ \mathbf{Y} = \mathbf{y} \\ \mathbf{Z} = \sin(\gamma)(\kappa^{-1} - R') \end{cases} \quad \text{with} \quad \begin{cases} r = sign(c)\cos(\beta)\sqrt{\mathbf{x}^2 + \mathbf{y}^2} \\ \beta = \arctan\frac{\mathbf{y}}{\mathbf{x}} \\ R' = \kappa^{-1} - \cos(\gamma)(\kappa^{-1} - r) \\ \gamma = \mathbf{z}/\kappa^{-1} \\ \kappa^{-1} = \frac{a_z}{|c|}; \end{cases}$$

where $-1 \leq c \leq 1$ is the bending control parameter, which, when zero, yields no bending (for $c = 0$ the deformation is not applied).

Finally a *circular* swelling deformation along the z axis is given by:

$$Swell(s,\ \mathbf{S}) = \begin{cases} \mathbf{X} = \mathbf{x} + sign(\mathbf{x})(R''\cos\alpha - (R'' - \sigma)) \\ \mathbf{Y} = \mathbf{y} + sign(\mathbf{y})(R''\cos\alpha - (R'' - \sigma)) \\ \mathbf{Z} = R''\sin\alpha \end{cases} \quad \text{with} \quad \begin{cases} \sigma = a_x s \\ R'' = (a_z^2 - \sigma^2)/(2\sigma) \\ \alpha = \arctan\frac{\mathbf{z}}{(\mathbf{R}' - \sigma)} \end{cases}$$

where s is the swelling control parameter (zero for no swelling).

Following the suggestion made by [Solina & Bajcsy 90], the above deformations are applied in the following order: first tapering, then swelling and finally bending.

Once deformed, the shape is roto-translated to simulate the change in viewpoint by applying in sequence pan (about z) and tilt (about x) rotations, orthographic projection ($Proj$) and finally rotation about the optical axis y and translation in the image plane (by P_x and P_z). The whole chain of transformations of the initial 3D shape $\mathbf{S}$ to its full projection onto the image plane z-x $\mathbf{S}'$ is:

$$\begin{aligned} \mathbf{S}' = \begin{bmatrix} \mathbf{x}' \\ \mathbf{z}' \end{bmatrix} = Trasl(P_x, P_z, Rot_y(\theta_{opt}, Proj(Rot_x(\theta_{tilt}, Rot_z(\theta_{pan}, \\ Bend(c, Swell(s, Taper(K_x, K_x, \mathbf{S})))))))) \end{aligned} \tag{6.2}$$

Now we are ready to describe the construction of the PDCM of geons. The knottiest problem is to determine the occluding contour. For doing this, the following approximation has been employed.

The transformation chain in Eqn. (6.2) is applied to the two bases of the superelliptical cylinder and then we take the four outermost points $P1'_a$, $P1'_b$ and $P2'_a$, $P2'_b$ (small circles in Fig. 6.1-right-B) and find the two corresponding points in the original undeformed superellipses (small circles in Fig. 6.1-right-A). These two pairs of points are

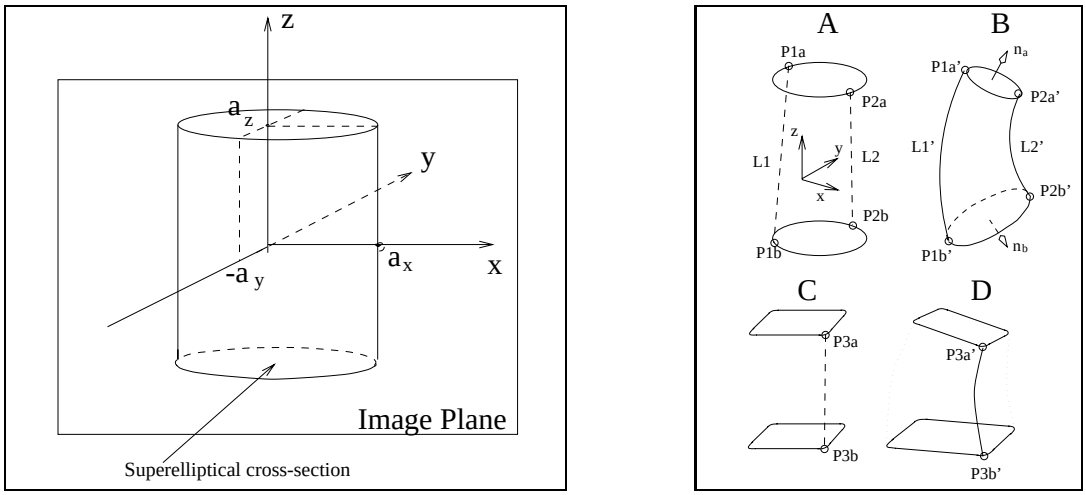


Figure 6.1: Construction of the parametrically deformable contour model of geons: Initial superelliptical cylinder (left) and determination of occluding contour and central rim (right). See text for details.

linked by two 3D straight lines L_1 and L_2, as shown in Fig. 6.1-right-A and are then deformed according to Eqn. (6.2) and the resulting $L1'$ and $L2'$ (Fig. 6.1-right-B) will be then used as the two sides of the occluding contour.

By checking the projection on the image plane of the normals' $\mathbf{n}_a$ and $\mathbf{n}_b$ to the superelliptical ends, it is possible to determine whether each of the two ends is visible or not: if visible, the whole superellipse contour will be added to the geon PDCM; otherwise only its outermost part between $P1'_a$ ($P1'_b$) and $P2'_a$ ($P2'_b$) will be included in the final PDCM.

In the case where the geon has a square cross-section (small ϵ, say less than 0.5 in the superelliptical cross-section model) the central edge is determined by joining the two corners $P3_a$ and $P3_b$ (Fig. 6.1-right-C) from the undeformed superelliptical bases occurring at $\eta = \pi/4$ in Eqn. (6.1) by a 3D straight line and then deforming it by Eqn. (6.2); the resulting 2D curve is shown in Fig. 6.1-right-D.

The PDCM described above is controlled by 12 parameters, namely a_x, a_y, a_z, ϵ, K_x, s, c, θ_{pan}, θ_{tilt}, θ_{opt}, P_x, and P_z. All these controlling parameters immediately relate to those of a DSQ, therefore they have a 3D meaning as we shall see in the experiments (in particular in Sec. 6.7.2 and 6.7.3) where a tridimensional rendering of the fitting

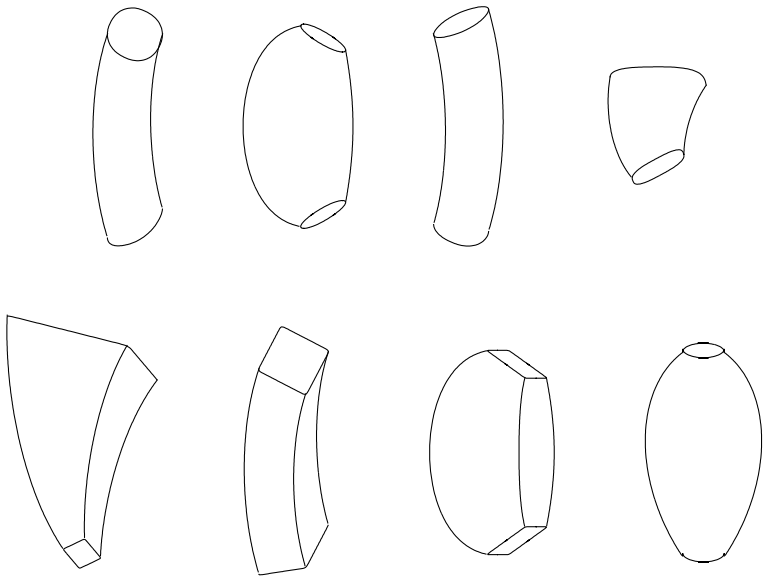


Figure 6.2: Examples of geon contour models generated by the proposed method. The parameters controlling the PDCM shape are the same as the ones that would produce a similar contour projection from a globally deformable superquadric.

results will be given.

By these simple approximated models of geon contour inspired by deformable superquadric modelling, we can represent 12 geon classes with a good level of accuracy. The proposed model could actually represent all 36 geon classes once a certain amount of deformation is introduced that would asymmetrically deform the superelliptical cross section; however this is unnecessary, because it has been shown that such deformations are unrecoverable from 2D images [Metaxas *et al.* 93].

Some examples of geon PDCMs produced by this method can be seen in Fig. 6.2 and in the experimental section. The time for creating an instance of such a model is less than 1ms on a SPARC 10 machine, which is over 2 orders of magnitudes faster that any other method that would use a direct exact computation of the outline using raster scan techniques or computation of surface normals as in [Metaxas *et al.* 93].

It is necessary to point out that, although effective, this model becomes rather imprecise with high amounts of bending under viewing directions where the tilt is greater than about $\pi/4$; in these situations, however, the geon would be virtually unrecoverable

from its contour, unless a precise model of it is known.

One last important remark is due. Geons are, by their very nature, qualitative primitives and one might ask how can they be modelled by simple shapes – such as the PDCM one proposed here – or by globally deformable superquadrics. Although this criticism is certainly correct, for the task of recognition and detection these models constitute a valid *low order* approximation of geon shapes and surely good enough to recover their distinguishing features. It is up to the fitting algorithm to be able to cope with this low order-ness and make sure that high-order components do not affect the robustness of the process.

6.4 Aspect partitioning of PDCM

In the previous subsection, a PDCM has been presented which represents the variable contour of geons through its parameters. This section describes how the PDCM parameter space is partitioned in "cells" that correspond to topologically distinct PDCM aspects.

First, the definition of topological equivalence for geon PDCMs is given, and that will be used to generate distinct aspects. Let us take the model described in the previous section and give it an orientation corresponding to the direction of the positive z axis of the original undeformed superelliptical cylinder.

Now, let us impose a labelling scheme on some features of the geon PDCM. Let $U = \{curved, squared\}$ be two properties of the cross section, and $V_{top} = \{visible/non\text{-}visible\}$ and $V_{bottom} = \{visible/non\_visible\}$ two properties of the two geon ends which indicate whether they are visible or not, the ends being the top and bottom superellipses in Fig. 6.1-left.

The Cartesian product $U \times V_{top} \times V_{bottom}$ produces 8 PDCM classes. Of the twelve PDCM parameters, only four change the PDCM class, namely ϵ, which affects the cross-section roundness, and c (bending), θ_{tilt} and θ_{pan}, which affect the visibility of the two ends. Cross-section dimensions, length, tapering and swelling do not change the topology as it has been defined. For the topology theory connoisseurs, these equivalence classes partition the 4D parameter space $S = \{\epsilon, c, \theta_{tilt}, \theta_{pan}\}$ into eight dense

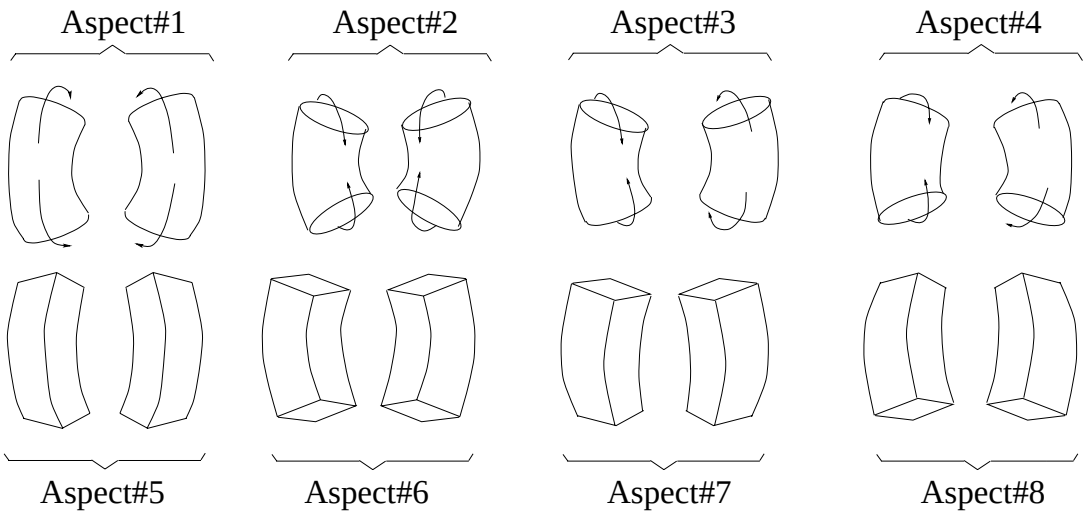


Figure 6.3: Distinct PDCM topologies and their enumeration. The features defining the topology are the visibility of top and bottom ends and the central rim.

simply-connected open subspaces of S, thus creating eight different topologies in the parameter space; each of these topologies correspond to a stable view of the PDCM that preserve the labelling we have imposed; these topologies are known as *aspects* [Koenderink & vanDoorn 79] of the PDCM, of which some examples are shown in Figure 6.3 along with the enumeration that will be used henceforth. The topology-constrained PDCM is rightfully called *parametrically deformable aspect* (PDA).

As said in the previous subsection, the property $U = \{curved, squared\}$ is determined by simply setting a threshold $\overline{\epsilon} \in 0.3 \,..\, 0.6$ for ϵ, hence dividing S in two symmetric 3D sub-spaces S' and S''.

The separation from one topology to another in S' (S'') are singularities that are called *visual events surfaces* [Koenderink & vanDoorn 79]. By analysing the expressions of the two normals to the ends as functions of c, θ_{tilt} and θ_{pan} from 6.3, a closed-form for those surfaces has been determined as the zero set of the functions A and B defined as follows:

$$\begin{cases} A = \cos(\theta_{tilt})\sin(\theta_{pan})\sin(\alpha) - \sin(\theta_{tilt})\cos(\alpha) \\ B = \cos(\theta_{tilt})\sin(\theta_{pan})\sin(\alpha) + \sin(\theta_{tilt})\cos(\alpha) \\ \alpha = \arctan(c) \end{cases}$$

The plot in Figure 6.4 shows these surfaces. The region within which each aspect is defined is given by the inequalities in the table of Figure 6.4.

Aspect#1:	$\epsilon > \overline{\epsilon}$	$A > 0$	$B > 0$
Aspect#2:	$\epsilon > \overline{\epsilon}$	$A < 0$	$B < 0$
Aspect#3:	$\epsilon > \overline{\epsilon}$	$A > 0$	$B < 0$
Aspect#4:	$\epsilon > \overline{\epsilon}$	$A < 0$	$B > 0$
Aspect#5:	$\epsilon \leq \overline{\epsilon}$	$A > 0$	$B > 0$
Aspect#6:	$\epsilon \leq \overline{\epsilon}$	$A < 0$	$B < 0$
Aspect#7:	$\epsilon \leq \overline{\epsilon}$	$A > 0$	$B < 0$
Aspect#8:	$\epsilon \leq \overline{\epsilon}$	$A < 0$	$B > 0$

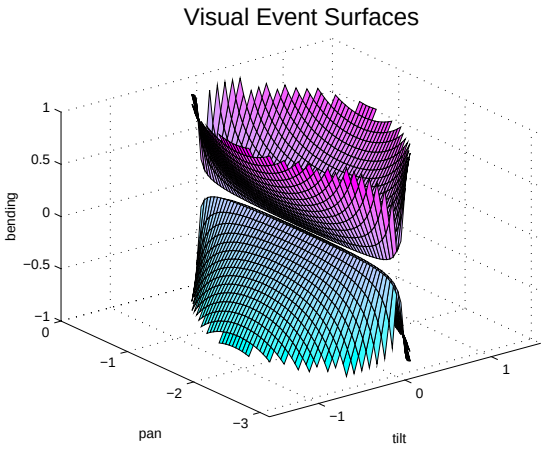


Figure 6.4: Aspect definition (left table, see text for the definitions) and plot of the visual event surfaces in the bending/pan/tilt parameter subspace (bottom-hull: Aspect#1/5; top-hull: Aspect#2/6; right-part: Aspect#3/7; left-part: Aspect#4/8). The gap between the hulls is a rendering flaw.

In principle it should be possible to consider also aspects without one or both ends to model parts that are joined to other parts at their ends. All the discussion so far and what follows can be trivially extended to include these other aspects.

6.5 Matching a single aspect

Our Model-Based Optimisation approach to geon recovery involves the minimisation of an objective (or cost) function that expresses the quality of the image-model match and other constraints that will be discussed later.

There are many conceptually different ways of designing an objective function suitable for a certain application but they mainly fall in these three categories: Energy Minimisation (EM), Maximum A Posteriori (MAP) or Minimum Description Length (MDL). It has been shown that given a certain problem and a certain fitting quality assessment criterion, they are conceptually equivalent (i.e. in [Zhu & Yuille 95, Leclerc 89]). Practically, however, the nature of a particular problem makes the use of a particular method easier. In this work, a MAP philosophy has been followed, but the ideas behind it could be restated in MDL terms.

Let $\mathcal{H}_i = \mathcal{H}(\mathbf{x}_i)$ be a geon PDCM instance built as in Sec. 6.3 expressed in terms of

pixels by a set of (i, j) image pixel coordinates and of which we would like to determine the likelihood of fit, and let

$$\mathbf{x_i} = [a_x \ a_y \ a_z \ \epsilon \ K_x \ s \ c \ \theta_{pan} \ \theta_{tilt} \ \theta_{opt} \ P_x \ P_z]^T \tag{6.3}$$

be the vector of the PDCM parameters. Furthermore, let $\mathcal{I}$ be the original image and $\mathcal{E}$ the binary edge image, which can be produced by a standard Canny edge detector; $\mathcal{E}$ has the same shape as $\mathcal{I}$ and $(i, j) \in \mathcal{E}$ is 1 if an edge has been detected at $(i, j) \in \mathcal{I}$ and 0 otherwise.

The *a posteriori* likelihood of a PDCM matching the image can be expressed in term of *a priori* probabilities by Bayes rule:

$$P(\mathcal{H}_i \mid \mathcal{E}) = \frac{P(\mathcal{E} \mid \mathcal{H}_i) \ P(\mathcal{H}_i)}{\sum_{j=1}^{N_h} P(\mathcal{E} \mid \mathcal{H}_j) \ P(\mathcal{H}_j)} \tag{6.4}$$

where N_h is the total number of hypotheses produced by the optimisation procedure.

The model that best fits the image is the one for which $P(\mathcal{H}_j \mid \mathcal{E})$ is maximum, that is:

$$\mathcal{H}_{best} = \mathcal{H}(\mathbf{x}_{best}) = \max_i \{P(\mathcal{H}_i \mid \mathcal{E})\}$$

or, by inverting the sign and expressing probability in term of logarithms:

$$\mathcal{H}_{best} = \mathcal{H}(\mathbf{x}_{best}) = \min_i \{-\log(P(\mathcal{H}_i \mid \mathcal{E}))\} \tag{6.5}$$

Since the denominator of Eqn. (6.4) is constant over all hypotheses, the minimisation need only be concerned with the numerator. In the two following sections we describe how we defined the model-conditional image and prior probabilities.

6.5.1 Model-conditional image probability

In Eqn. (6.4) $P(\mathcal{E} \mid \mathcal{H}_i)$ expresses the conditional probability of having particular image evidence in the presence of the model. Although many ways of defining this probability are possible, here this probability is expressed in terms of how many image edgels "match" the PDCM contour.

Let

$$\mathcal{E}_m(\mathcal{H}_i) = \{(k, l) \ : \ \mid (i, j) - (k, l) \mid \leq d \ , \quad (i, j) \in \mathcal{H}_i\}$$

be the d-neighbourhood of the model contour $\mathcal{H}_i$ and $\mathcal{E}_b(\mathcal{H}_i) = \mathcal{E} - \mathcal{E}_m(\mathcal{H}_i)$ the rest of the edge image which is not covered by it; henceforth we drop the $\mathcal{H}_i$ arguments wherever there cannot be ambiguities.

By assuming that the presence/absence of an edge in $\mathcal{E}_b$ and $\mathcal{E}_m$ can be considered independent (this is valid in general) and with different distributions, $P(\mathcal{E} \mid \mathcal{H}_i)$ can be expressed as:

$$P(\mathcal{E} \mid \mathcal{H}_i) = P(\mathcal{E}_b \mid \mathcal{H}_i) \cdot P(\mathcal{E}_m \mid \mathcal{H}_i). \tag{6.6}$$

$\mathcal{E}_b$ and $\mathcal{E}_m$ can be considered, to a first approximation, as realizations of binary ergodic processes, for which the probability of single local outcomes are all the same, namely p_{b1} and p_{m1}, respectively. The value of p_{b1} is given by the ratio between edge locations and the number of pixels in the image (typical values: 0.02-0.06) and p_{m1} ranges from 0.6 to 0.9, depending on the neighbourhood dimension d and how good the edge detection is expected to be. This ergodicity assumption is simplistic; a Markovian model that would take into account relationships between neighbouring pixels would perhaps be a more accurate model but this is left for future work.

Let N_{b0}, N_{b1}, N_{m0} and N_{m1} be the number of locations (i, j) that are "1" (edge) or "0" (non-edge) in $\mathcal{E}_b$ and $\mathcal{E}_m$, respectively; the probability that a certain number of elements in $\mathcal{E}_b$ and $\mathcal{E}_m$ is "1" or "0" follows a binomial distribution but, since we are interested in a particular realization of the process that is the image itself, the two probabilities in Eqn. (6.6) can be expressed as:

$$P(\mathcal{E}_b \mid \mathcal{H}_i) = p_{b1}^{N_{b1}}(1 - p_{b1})^{N_{b0}}$$
$$P(\mathcal{E}_m \mid \mathcal{H}_i) = p_{m1}^{N_{m1}}(1 - p_{m1})^{N_{m0}}$$

By taking the logarithm of both sides, we obtain:

$$\log(P(\mathcal{E}_b \mid \mathcal{H}_i)) = N_{b1}\log(p_{b1}) + N_{b0}\log(1 - p_{b1}))$$
$$\log(P(\mathcal{E}_m \mid \mathcal{H}_i)) = N_{m1}\log(p_{m1}) + N_{m0}\log(1 - p_{m1}))$$

which in turn, by letting $N_1 \triangleq (N_{b1} + N_{m1})$ be the overall number of pixels in the image that are edge, are expanded to:

$$\begin{aligned} \log(P(\mathcal{E}_b \mid \mathcal{H}_i)) = [N_1\log(p_{b1}) + N_1\log(1 - p_{b1})] - \\ (N_{m1}\log(p_{b1}) + N_{m0}\log(1 - p_{b1})) \\ \log(P(\mathcal{E}_m \mid \mathcal{H}_i)) = N_{m1}\log(p_{m1}) + N_{m0}\log(1 - p_{m1}). \end{aligned} \tag{6.7}$$

Then by taking the logarithm of both sides of Eqn. (6.6) and expanding we obtain:

$$\log(P(\mathcal{E} \mid \mathcal{H}_i)) = \log(P(\mathcal{E}_b \mid \mathcal{H}_i)) + \log(P(\mathcal{E}_m \mid \mathcal{H}_i)) =$$
$$K + [N_{m1}\log(p_{m1}) + N_{m0}\log(1-p_{m1})] - [N_{m1}\log(p_{b1}) + N_{m0}\log(1-p_{b1})] \quad , \tag{6.8}$$

where K is the constant term in square brackets in Eqn. (6.7) and therefore it will be dropped in the MAP estimation.

In an information theoretical framework this equation has a precise meaning. The term $-\log(P(\mathcal{E} \mid \mathcal{H}_i))$ is the overall number of bits necessary to express the whole edge image $\mathcal{E}$ and $-\log(P(\mathcal{E}_m \mid \mathcal{H}_i))$ and $-\log(P(\mathcal{E}_b \mid \mathcal{H}_i))$ are the number of bits needed to represent the information in the model neighbourhood ($\mathcal{E}_m$) and in the background ($\mathcal{E}_b$) under the ergodicity assumption. The minimisation in Eqn. (6.5) can then be re-interpreted as the search for the most economical description in term of the edge evidence *and* the model, bringing all into a MDL framework [Pednault 89, Leclerc 89]. A more formal proof of the MDL/MAP equivalence can be found in [Rissnanen 83] and in the context of computer vision in [Fua & Hanson 89, Leclerc 89]. This information theoretical avenue was followed in [Fua 89] but with the fundamental difference that there the *a priori* p_m was computed by looking at the number of pixels matching the *current* instance of the model, therefore making the mistake of using the same data set for both training and estimation; some experiments that I carried out by using their objective function gave unusually high likelihoods for bad fits as well, which was somehow expected from what has been just said.

Fig. 6.5 shows an example behaviour of $-\log(P(\mathcal{E} \mid \mathcal{H}_i))$ (up to K) for $p_{m1} = 0.7$, $p_{b1} = 0.06$ and the total number of model points $N_m = N_{m1} + N_{m0}$ ranging from 100 to 300; the small step has been added in order to show the points at which the absence or presence of the model $\mathcal{H}_i$ is equally likely ($P(\mathcal{E} \mid \mathcal{M}_i) = 0.5$): beyond this line the probability increases with the overall model dimension N_m, that is a preference is given to bigger models.

6.5.2 Model prior probability: A heuristic

Within a Bayesian framework it is necessary to express the occurrence probability of each instance of the model, called the *model a priori probability*. In most research this

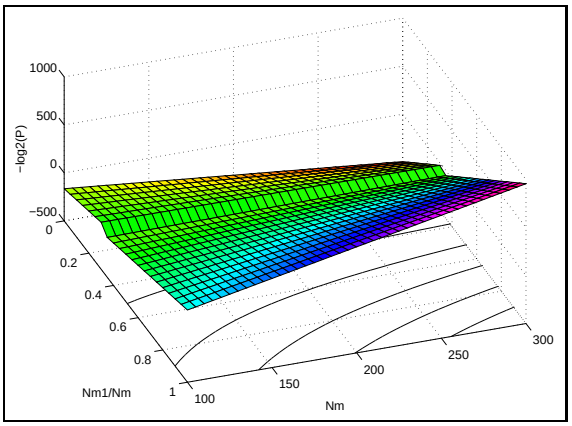


Figure 6.5: Example of model-conditional image probability $-\log(P(\mathcal{E} \mid \mathcal{H}_i))$ for $p_{m1} = 0.7$, $p_{e1} = 0.06$. See text for details.

probability is neglected (i.e. is considered uniform) but, through experimentation, it has been found that by introducing a heuristic on the prior probabilities, the overall quality of the fitting can be improved.

The reasons for introducing a model prior probability are essentially three: *i*) some parameter configurations are unlikely to occur (such as a bent and swollen object); *ii*) certain configurations of parameters arise from a weird viewpoint that would make detection impossible; and *iii*) it biases the fitting to more perceptually likely shapes. These considerations are both practical and also correspond to sensible assumptions to reduce the quantitative shape ambiguities caused by the projection.

A sensible heuristic has been defined to express these loose constraints. The probability of each aspect is expressed by overlapping (multiplying) marginal densities of parameter values or combinations of them, tacitly assuming independence amongst them. The parameters we took into consideration are the dimension parameters a_x, a_y and a_z, swelling, bending and the pan rotation; the others are given a uniform probability. Below, the definition of the probability density functions is given.

c and θ_{pan} In the case of Aspects #3, #4, #7 and #8 when θ_{pan} is close to $-\pi/2$, (when the viewing direction is parallel to the bending plane) the bending cannot be detected from the occluding contour alone and therefore we need to strongly assume straightness of axis, i.e. the only thing we can perceive in these situations.

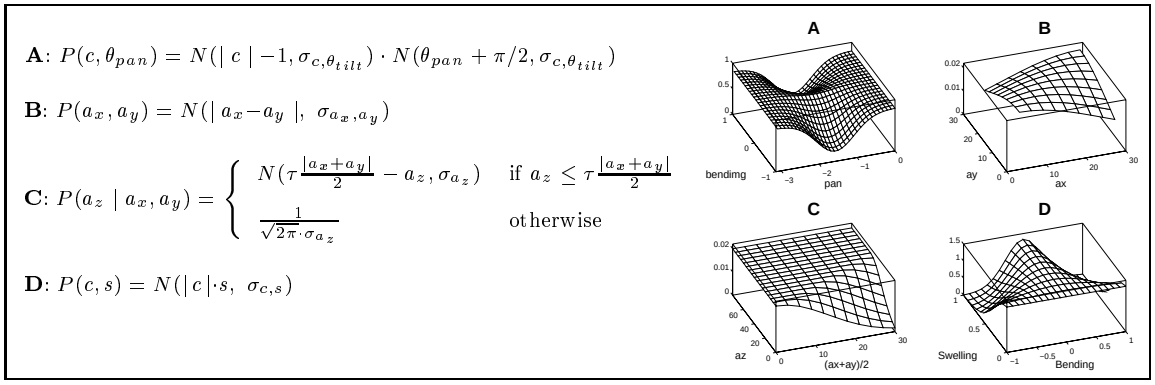


Figure 6.6: Heuristic model prior probabilities: definitions and plot for each contributing term. The definitions and details are given in the text. These probabilities constitute an heuristic that bias the fitting to perceptually more plausible volumetric shapes corresponding to similar 2D contour projections.

Without this constraint the model could bend forward an arbitrary amount and yield essentially the same occluding contour. To model this constraint we set up an unnormalised p.d.f. like the one Fig. 6.6-A, shown for $\sigma_{c,\theta_{tilt}} = 0.5$. In fitting Aspects #1, #2, #5 and #6, the bending is essential for the visibility or invisibility of both ends and this constraint is not used.

a_x and a_y The projection onto the image plane of a 3D object changes its shape, but our perceptual system is slightly biased to assume more compact cross-sections rather than weird rotation angles [Lowe 85]. We therefore model the joint p.d.f. as given in Fig. 6.6-B, which is a constant-height ridge running along the $a_x = a_y$ line. The value of σ_{a_x,a_y} is fairly large because this constraint need not be severe ($\sigma_{a_x,a_y} = 20$ in Fig. 6.6-B). This constraint assumes that the objects in the scene are not too flat and should be dropped if that is the case.

a_z The PDA length could take any value but, since a_z defines the length of allegedly elongated parts like geons, it should be biased to be bigger than the cross-section dimensions by a constant factor τ. A non-normalised p.d.f. given in Fig. 6.6-C has been set to model this constraint; the figure shows it for $\tau = 1.5$ and $\sigma_{a_z} = 20$.

c and s High swelling and bending are incompatible. In statistical terms we can express this constraint by a (non-normalised) p.d.f. like the one shown in Fig. 6.6-D and arising from a Gaussian distribution over the product $c \cdot s$. The plot in Fig. 6.6-D is given for $\sigma_{c,s} = 0.3$.

Now that we have all the non-normalised probabilities and given the assumption of prior independence between parameters, we just multiply them together to obtain the (non-normalised) *a priori* p.d.f. of the model:

$$\begin{aligned}\log(P(\mathcal{H}_i)) = \quad & H + \log(P(a_z \mid a_x, a_y)) + \log(P(c, s)) + \\ & \log(P(a_x, a_y)) + \log(P(c, \theta_{pan})).\end{aligned} \tag{6.9}$$

The normalisation constant H is unnecessary because it does not affect the MAP estimate.

This heuristic has improved the perceptual goodness of the recovered shapes but there would be other possible ways of defining the model prior probability, which could also incorporate more detailed specific domain-dependent knowledge about the scene structure.

6.5.3 Maximum a Posteriori estimation

The MAP estimation obtained by the minimisation of

$$-\log(P(\mathcal{H}_i \mid \mathcal{E})) = -\log(P(\mathcal{E} \mid \mathcal{H}_i)) - \log(P(\mathcal{H}_i)), \tag{6.10}$$

where the two terms are given by Equations (6.9) and (6.8), is rather difficult to achieve, since it is extremely irregular and presents many shallow and/or narrow minima.

As an example, Figure 6.7 shows some graphs of the objective function value taken at three orthogonal planar regions of the parameter space (in particular about the initial estimate of the handset upper-piece example of Figure 6.10): although the three surfaces are rather rugged, three pronounced valleys stand out that correspond to good values of the objective function. In the middle figure, however, two valleys beyond the ripples might jeopardise the fitting procedure.

By trying to minimise Eqn. (6.10) alone, it was also found that sometimes the optimisation got stuck in local minima because of the step-like nature of the model-conditional probability of Eqn. (6.8) (remember we used a binary "belonging to the model" criteria). For improving this situation (but see Section 6.7.4), a small smoothing term has been added to the right side of Eqn. (6.10); this term represents the average minimal distance between contour model and image edge points (by using a minimal distance

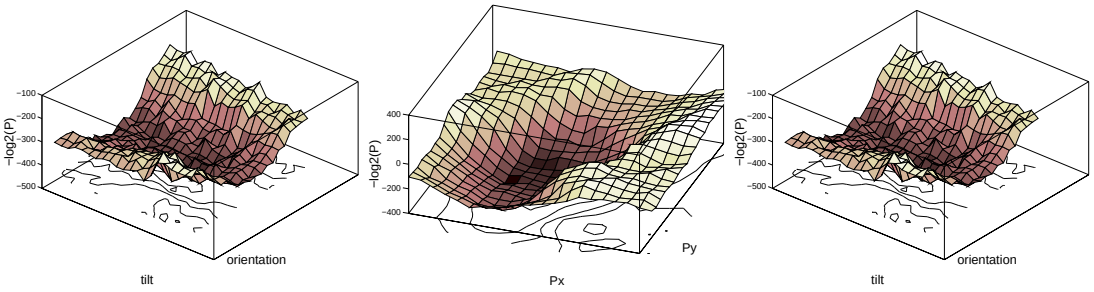


Figure 6.7: Three graphs of the objective function value taken at three orthogonal planar regions of the parameter space about the initial estimate of the handset upper-piece example of Figure 6.10: although the three surfaces are rather rugged, three pronounced valleys stand out that correspond to good values of the objective function.

transform computed off-line) and it does not affect the MAP estimate but just helps convergence in cases where image and model are so much displaced that the objective function does not change due to the low number of edge points falling inside the model neighbourhood, thereby making optimisation troublesome. This term can then be seen as "telling the optimisation where to go" in absence of other information.

In early stages of the work, a Levenberg-Marquandt method with added random perturbations was used, following [Borges 96] and other works, but this method led to difficult convergence. The choice fell then to Simulated Annealing (see Appendix C for a summary of the method), which is a powerful optimisation tool that efficiently combines gradient descent and controlled random perturbations to perform the minimisation of non-convex functions. The actual implementation is a publicly available version of Simulated Annealing, called Adaptive Simulated Annealing (ASA) [Les93]. The set-up of the ASA algorithm will be extensively discussed in the next section.

6.6 Experimental system

This section outlines the simple experimental system, schematically depicted in Fig. 6.8, that has been used to carry out the experiments.

Starting from the set of hypotheses produced by the method described in previous chapters, for each hypothesis, each of the eight PDA is initialised at a representative position and independently fitted to the image. The PDA that obtains the best scores

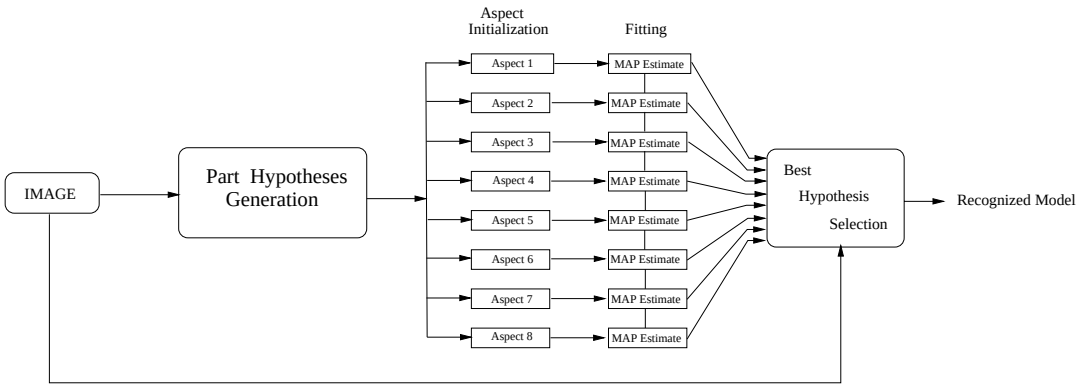


Figure 6.8: The simple aspect-based control strategy. For each part hypothesis, the eight PDAs are independently initialised and fitted to the image. The one that obtains the best fitting score gives the best interpretation of the image.

is considered the best fit to the image.

The approach relies on two fundamental assumptions [Eggert *et al.* 95]:

1. The MAP estimate that started with the "correct" hypothesis will converge to the correct interpretation of the image;
2. The quality of the fit (score) of this correct interpretation must be higher than any other.

No theoretical proof of convergence and uniqueness of the method is possible since the problem is strongly non-linear and too complex to be analysed (as stated also by [Eggert *et al.* 95], where rigid models were used). The experiments of the next section will, however, empirically show that the proposed method reasonably complies with these two goals.

In the following three subsections, the PDA initialisation and the optimisation set-up of the experimental system is described.

6.6.1 Initialisation

The initialisation stage is concerned with estimating coarse part initial hypotheses (sometimes called the *frame* [Subirana-Vilanova 93]) that comprise position, orientation of the major axis and dimensions. These initialisations need not be precise and the degree of allowed inaccuracy depends upon the power of the optimisation procedure. The initial estimates are produced by the part-grouping and filtering method proposed in previous chapters. However, the two modules are currently not integrated and in some experiments the initialisations have been set by hand to be qualitatively similar to those output by the MDL hypothesis filtering method of Chapter 5.

It is worth highlighting again that the instantiation of deformable models to edge images is a relative novelty in vision, functionally matched only by the system proposed by [Metaxas *et al.* 93]. Part hypotheses are physically PDM models (Sec. 3.4) fitted onto the image and their position, orientation and respective variation modes are used to initialise the PDCM by assigning them to $\overline{P}_x$, $\overline{P}_z$, $\overline{a}_z$, $\overline{a}_x = \overline{a}_y$ and $\overline{\theta}_{opt}$, respectively. The values of $\overline{a}_x$ and $\overline{a}_y$ are set to be equal because we do not have prior information on the aspect-ratio of the part cross-section. Information about bending, tapering or roundness, which are projective quasi-invariant properties, are not currently used to better initialise PDCMs, but this could be done with relatively little effort.

6.6.2 Aspect hypotheses generation

From each of the initial frames, all eight distinct aspects are instantiated by properly setting the PDCM parameters that control the aspect topology. Referring to Fig. 6.4, I chose points in the topology-controlling parameter space (Sec. 6.4) that are more or less equidistant from the visual event surfaces and therefore are placed in a fairly central position within each aspect cell. This choice is a sensible heuristic that reduces the distance between the initial point and any possible true final estimate[4]. Table 6.1 (along the "init" columns) shows these values for each aspect topology. The other parameters, bending, swelling and tapering, were all set to zero.

[4] The choice of these values can be regarded also as giving *maximal disambiguation distance* between visual events [Kender & Freudenstain 87].

	squareness (ϵ)			bending (c)			θ_{tilt}			θ_{pan}		
	init	min	max	init	min	max	init	min	max	init	min	max
Aspect#1:	0.75	0.51	0.99	-0.5	-1	0	0	$-\pi/4$	$\pi/4$	$-\pi/2$	$-\pi$	0
Aspect#2:	0.75	0.51	0.99	0.5	0	1	0	$-\pi/4$	$\pi/4$	$-\pi/2$	$-\pi$	0
Aspect#3:	0.75	0.51	0.99	0	-1	1	$-\pi/8$	$-\pi/4$	0	$-\pi/2$	$-\pi$	0
Aspect#4:	0.75	0.51	0.99	0	-1	1	$\pi/8$	0	$\pi/4$	$-\pi/2$	$-\pi$	0
Aspect#5:	0.25	0.05	0.49	-0.5	-1	0	0	$-\pi/4$	$\pi/4$	$-\pi/2$	$-\pi$	0
Aspect#6:	0.25	0.05	0.49	0.5	0	1	0	$-\pi/4$	$\pi/4$	$-\pi/2$	$-\pi$	0
Aspect#7:	0.25	0.05	0.49	0	-1	1	$-\pi/8$	$-\pi/4$	0	$-\pi/2$	$-\pi$	0
Aspect#8:	0.25	0.05	0.49	0	-1	1	$\pi/8$	0	$\pi/4$	$-\pi/2$	$-\pi$	0

Table 6.1: Initialisation and bounds for the aspect topology-controlling parameters. See text for details.

6.6.3 Optimisation set-up

The optimisation of strongly non-linear functions is *"typically a non-typical problem"* [Rao 84] and therefore no canned optimiser can be used. As pointed out in [Les93], the set-up of the ASA algorithm is a bit tricky, since no theoretical guide exists, but once the right configuration has been found, the method becomes reasonably robust. Having said that, here we describe the essential set-up of the ASA optimiser.

One of the key decisions when using a constrained optimisation algorithm is the choice of the parameter bounds; the ASA algorithm requires hyper-rectangular bounds defined by a minimum and a maximum for each parameter.

Within our aspect-based control strategy, we basically have two sets of parameters, those controlling the PDCM aspect topology (ϵ, c, θ_{tilt} and θ_{pan}) and those that do not change it (a_x, a_y, a_z, K_x, s, P_x P_z and θ_{opt}).

Section 6.3 gave closed-form expression of the visual event surfaces bounding different aspect topologies. In order to make the ASA optimiser "stay within" a certain aspect topology, we do two things: *(1)* give it 4D search bounds (given in Table 6.1) that enclose the true aspect cell; and *(2)* invalidate states (through a specific ASA option) that fall outside the chosen aspect cell by checking the constraints given in the table in Fig. 6.4. In most of the experiments we carried out, the ratio between invalid and valid generated states was always less than 5%.

Besides the parameters constraining the aspect topology, the others need bounds too.

Bounds for the tapering and swelling deformations are set to their full range (Sec. 6.3); in the case of position, sizes and orientations, bounds are set as tolerances with respect to the initial values $\overline{P}_x$, $\overline{P}_z$, $\overline{a}_z$, $\overline{a}_x = \overline{a}_y$ and $\overline{\theta}_{opt}$. The following table summarises these bounds (N is the resolution of the image):

	a_x	a_y	a_z	K_x	s	θ_{opt}	P_x	P_z
Min	$\overline{a}_x - 40\%$	$\overline{a}_y - 40\%$	$\overline{a}_z - 40\%$	-1.0	0.0	$\overline{\theta}_{opt} - \pi/8$	$\overline{P}_x - \frac{N}{10}\%$	$\overline{P}_z - \frac{N}{10}\%$
Max	$\overline{a}_x + 40\%$	$\overline{a}_y + 40\%$	$\overline{a}_z + 40\%$	1.0	1.0	$\overline{\theta}_{opt} + \pi/8$	$\overline{P}_x + \frac{N}{10}\%$	$\overline{P}_z + \frac{N}{10}\%$

In order to improve convergence, we need also to specify the deltas for computing the partial pseudo-derivatives of the cost function, which are chosen such that for each parameter, a perturbation equal to its respective delta should produce detectable changes in the image at a given resolution. For 128x128 images, the values of Δ_{a_x},Δ_{a_y},Δ_{a_z}, Δ_ϵ, Δ_{K_x}, Δ_s, Δ_c, $\Delta_{\theta_{pan}}$, $\Delta_{\theta_{tilt}}$, $\Delta_{\theta_{opt}}$, Δ_{P_x} and Δ_{P_z} are set to 1.0, 1.0, 1.0, 0.05, 0.2, 0.05, 0.05, 0.01, 0.01, 0.01, 1.0 and 1.0, respectively.

The annealing schedule plays an important role. We have experimentally found that sub-optimal schedules are also related to the aspect topology we are trying to fit, probably because of the different kind and number of features. For good convergence the Temperature_Ratio_Scale parameter [Les93] has been set to 10^{-12} for Aspect#1 ... Aspect#4 and to 10^{-10} for Aspect#5 ... Aspect#8. Finally the number of iterations has been set to 2000, which I found to be a good trade-off between speed (about 10s for each optimization run on a SPARC 10) and convergence; moreover, for the experiments carried out with 128x128 images, we set p_{b1}, p_{m1} and d (see Sec. 6.5.1) to 0.07, 0.85 and 1, respectively.

6.7 Experimental results

In this section, four sets of experiments are discussed. In the first set, several fitting experiments of geon PDCM are shown for both synthetic and real images with the purpose of verifying the validity of the cost function and the optimization. The second set aims at assessing the validity of the two assumptions in the use of aspects given in Section 6.6. In the third set, three fitting experiments to the familiar handset test image are given along with interpretation of the results; in particular, an example of what can happen when the aspect-based strategy in not used is also supplied. Finally, the robustness of the fitting is assessed in Section 6.7.4.

6.7.1 Testing the MAP fitting

In this subsection a number of single fitting experiments are shown that help assess the validity of the cost function and the optimization method for fitting the PDCM proposed in this chapter. Here the aspect-based strategy is not used but, but in some experiments some of the topology-definining parameters have been constrained, as we shall see later.

The experiments presented here can in turn be divided in two sets, which are described in the following. In both experiments, the initialization is performed manually and is intentionally set to be poor to test for worst cases.

FIRST SET

The set of 18 fitting experiments shown Fig. 6.9 was designed to assess convergence and viability of the cost function and the optimization procedure. Six geon-like objects were created with some plasticine and an image of them was then taken with a resolution of 512x512 pixels. A Canny edge detector was applied and the resulting cluttered edge image is shown in Image C (left) of Fig. 6.9. This image has been intentionally used without any post-processing – like cleaning and linking – because we wanted to test the convergence in hard conditions.

Afterwards, two synthetic images mimicking the original one were created, one with roundish primitives (Image A of Fig. 6.9) and the other one with squared cross-

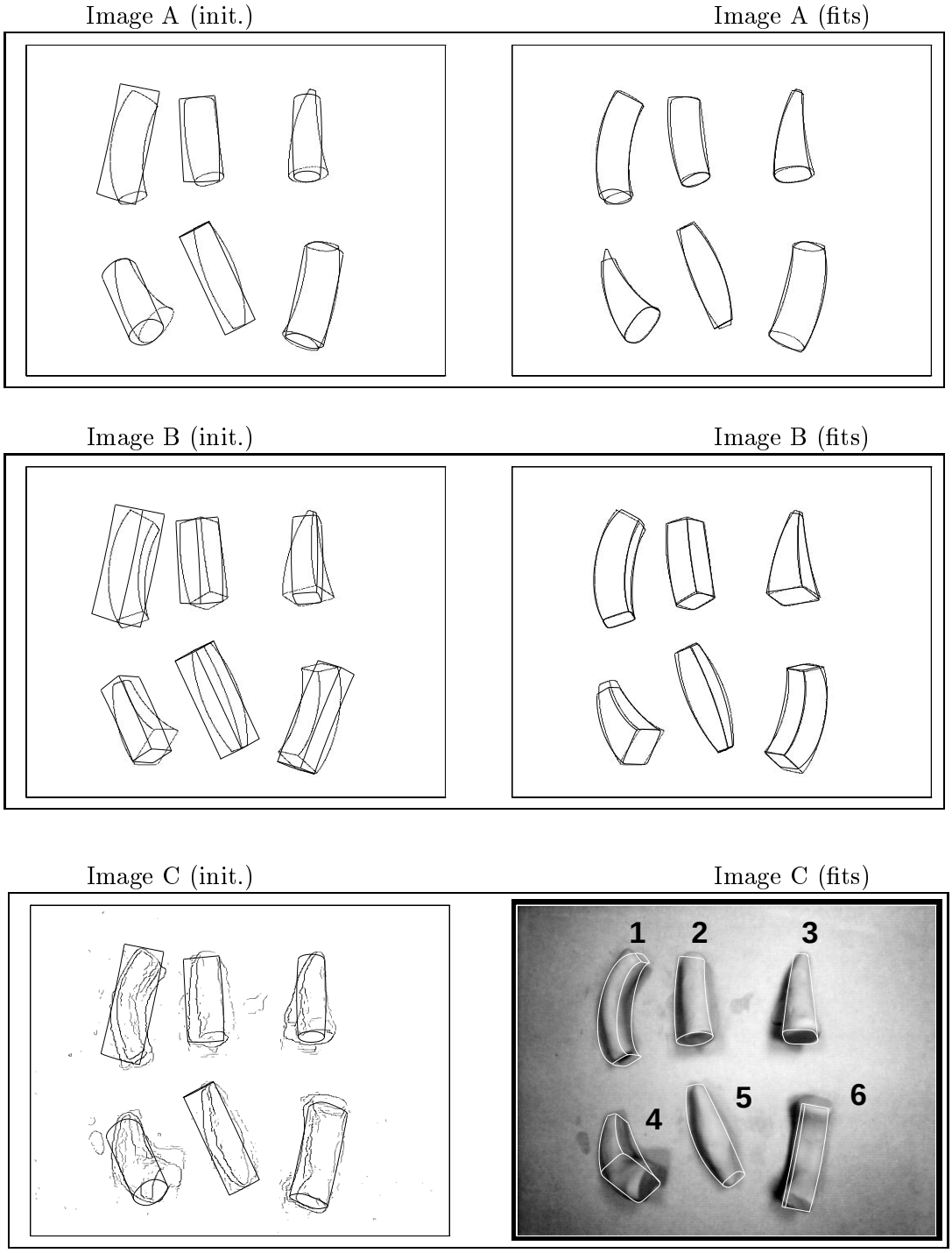


Figure 6.9: First set of experiments. The purpose is to assess validity of the objective functon and the optimization; the aspect-based strategy is not used here. A description of the eighteen fitting experiments is given in the text. Although only one initialization for each is shown here, many others have been tried that, however, kept the same initial topology as the ones shown.

sections (Image B of Fig. 6.9). The initial PDAs are shown in the left column Fig. 6.9 overlapped to the respective edge images; the initializations for these two synthetic images are rather crude but the right topologic aspects have been imposed to each example.

The manual initializations are the same across the three images except for the roundness parameter, which has been set to "squared cross-section" in image B. The corresponding results of the fitting can be seen in the right column. The neighbourhood dimension was set to 7 (that is $d = 3$) and the other parameters are the same as given in Sec. 6.6.3; each estimate was produced in about 25 seconds on a networked SPARC 10 machine.

- **Image A (Top of Fig. 6.9)** The results here are essentially good but in the case of Object 6 the sign of the bending is wrong. All the geon distinquishing features have been correctly detected, as can be visually seen.

- **Image B (Centre of Fig. 6.9)** In this case the results are better than the one in Image A because of the presence of the additional interior edge gives "more information" to the fitting.

- **Image C (Bottom of Fig. 6.9)** As expected, the results here are not particularly exciting but they can be considered positive, given the intentionally poor edge image quality we have used. Here, the roundness parameter ϵ was set free to check whether a change in the aspect topology would occour. The results for object 2,3 and 5 are very good. The fit of Object 1 is essentially correct (apart from slight tapering), but the spurious edge due to a high shading gradient caused the object to be interpreted as a bent prism. Object 4 too has been fitted rather poorly (because the high noise) but the essential orientation, bending and tapering have been recovered. In the case of Object 6 the presence of shadows and poor image contrast has been fatal and the fitting is a complete failure, with a final result that, although obtaining a higher score, looks poorer that the initial estimate.

	Handset Image			Banana&Mug	
	Top Piece	Handle	Bottom Piece	Banana	Cup
Tapering (K_x)	0.09	0.08	0.21	-0.02	-
Swell (s)	0.08	0.28	0.42	0.47	-
Bending (c)	-0.12	0.25	0.15	0.35	-
Squareness (ϵ)	0.84	0.26	0.69	0.45	-

Table 6.2: Final Parameter estimation. The recovered parameters allow a coarse description of the shape: Top Piece: cylinder; Handle: slightly bent prism; Bottom Piece: swollen and slighlty tapered cylinder; Banana: bent and swollen prism. See text for details.

SECOND SET

This set of experiments has been carried out with real images of isolated objects, a handset, a mug and a banana. The examples in Fig. 6.10 are 128x128 gray-level images; the neighbourhood dimension was set to 3 (that is $d = 1$) and the optimization set-up was the same as for the first set of experiments;

This time, the initialization was performed by manually selecting out rectangular regions of the image (top of Fig. 6.10), thresholding to extract the silhouette and finally by computing the principal moments that gave coarse estimates of position, axes lengths and orientation; the result are the initializations shown at the top of Fig. 6.10.

- **Handset** The top-left of Fig. 6.10 shows the original handset image with the initial models instances and their major axes overlapped on it. The two end parts (ear and mouth piece) have a rather poor initial estimate because of their low eccentricity and the shadows cast on the background. On the other hand, the central part is well defined and hence a good initial estimate is achieved; at this point there is no knowledge about the squareness of this part. The centre-left figure shows the edge image. It can be noticed that there is some cluttering, like that caused by circular ridges at the mouth piece. The bottom-left figure shows the results obtained after applying the optimisation to each one of the initial estimates. As it can be seen, the results are rather good. Table 6.2 shows that the main geons' distinguishing features are captured, with the exception of the top part (ear piece) not being swollen as it should; in this case, however, even for a human it would be difficult to tell the exact shape of such a short

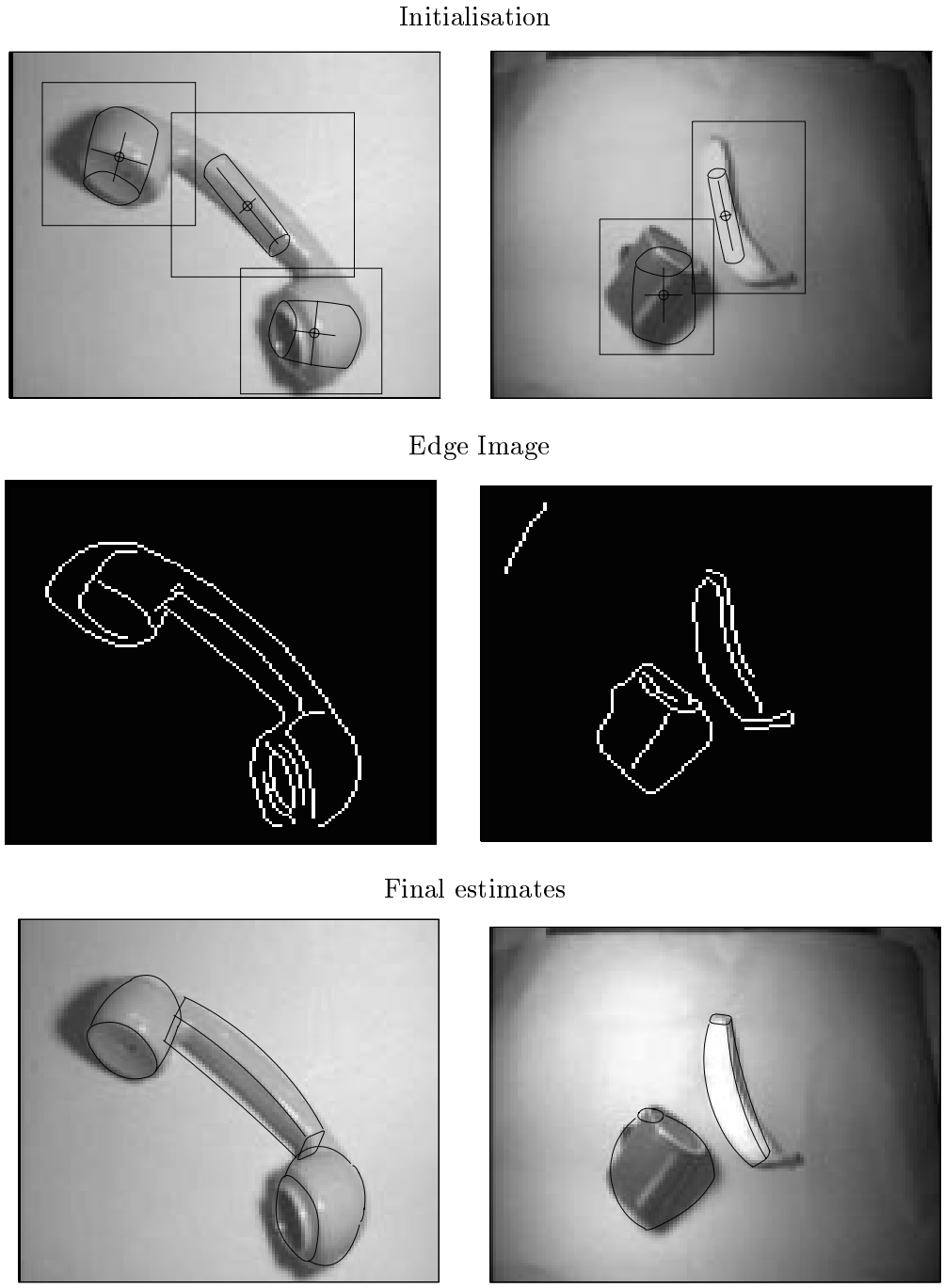


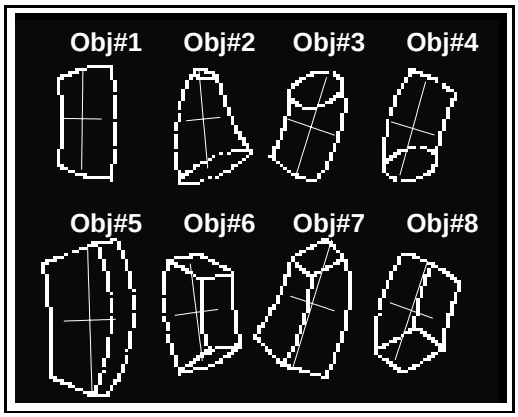
Figure 6.10: Second set of experiments with semi-automatic initialization again without using the aspect-based strategy (see text for details). The fitting to the handset geons and the banana are reasonably good whereas mug one is a sheer disaster.

part just from that poor edge image. Another remark worth making is that the length of the central part was correctly found despite the rightmost edge that runs along the whole handset. Note that some research has been recently carried out [Raja & Jain 92b, Borges 96] in the classification of geons from parameters such as those that define our PDCM.

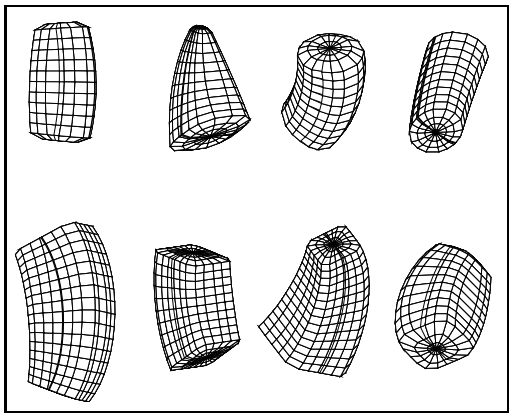
- **Banana** The top-right of Fig. 6.10 shows the initial estimate of the banana shape. The combined effect of a shadow in the right-hand side of the banana and poor resolution has lead to the poor edge image shown in the centre-right image. Here, the little incomplete square that somehow appears at the top and the double edge running along the right-hand side were interpreted as part of the shape, as shown in the final estimate in the bottom-right image. Table 6.2 shows that again all the essential features (apart from roundness, as just said) are grasped, such as curvature, swelling and no tapering.

- **Mug** This experiment is a complete failure. The big shadow, the highlight at the top and poor resolution led to an edge image that is virtually uninterpretable by the human eye. The initial estimate shown at the top-right of Fig. 6.10 is mis-oriented in the image plane by roughly $\pi/4$ and the estimation procedure produced a very poor result correspoinding to a deep local minimum of the objective function. Only by giving a very good initial estimate, a better rqesult was achieved.

The experiments described above show that the proposed method works reasonably well. The tests with clean images indicate that the optimization converges well. Results with real images show that the method performs well if a coarse initial estimate is given and there is not too much noise or spurious edges. However, as in the mug example, more care must be taken in determining the initial estimate because of if it is poor it can yield dramatically wrong results, especially for low-eccentricity objects and with high noise level.



A



B

	Aspect#1	Aspect#2	Aspect#3	Aspect#4	Aspect#5	Aspect#6	Aspect#7	Aspect#8
Obj#1	**-264.77**	-210.16	-170.32	-238.72	-211.14	-121.30	-133.60	-123.24
Obj#2	-242.49	**-314.16**	-252.11	-252.83	-167.71	-170.09	-106.96	-179.19
Obj#3	-177.34	-142.54	**-255.22**	-193.46	-183.00	-116.05	-13.18	-161.31
Obj#4	-236.55	-220.50	-187.26	**-261.39**	-177.06	-151.19	-166.85	-130.36
Obj#5	-252.05	-291.04	-245.37	-284.76	**-478.75**	-171.62	-62.20	-161.26
Obj#6	-290.78	-384.35	-346.86	-295.48	-392.26	**-458.83**	-362.63	-261.09
Obj#7	-284.68	-192.94	-300.22	-135.86	-166.14	-245.18	**-437.68**	-230.28
Obj#8	-249.61	-275.68	-211.26	-322.67	-270.43	-326.21	-241.54	**-374.05**

Figure 6.11: Experiment with synthetic images of 8 different aspects of geons and the confusion matrix representing the results of the fittings. The boxed results are the highest scoring PDA for each fitting experiment and all correspond to the PDA with the same topology as the respective test contours in fig. A. The superquadric corresponding to these best PDAs are displayed in figure B: the 3D shapes are well in agreement with the 3D structure that pops up from the contour images when we see them.

6.7.2 Testing the aspect-based strategy: Synthetic image

This subsection presents one of the experiments set up for testing the aspect-based strategy, in particular the two premises given at the beginning of Section 6.6: when starting from the correct PDA, the fitting must both converge and give a better score than the ones obtained from initialization with any of the wrong-topology aspects.

Eight synthetic contours of geons (Obj#1 ... Obj#8), each representing a different apect topology (Aspect#1 ... Aspect#8), have been placed in the same 128x128 image (Fig. 6.11-A) and a coarse initialization was given using estimates of just orientation, position, length and cross-section dimension; the initializations are represented by the crosses. Then, all eight distinct PDAs were initialized by the method given in Section

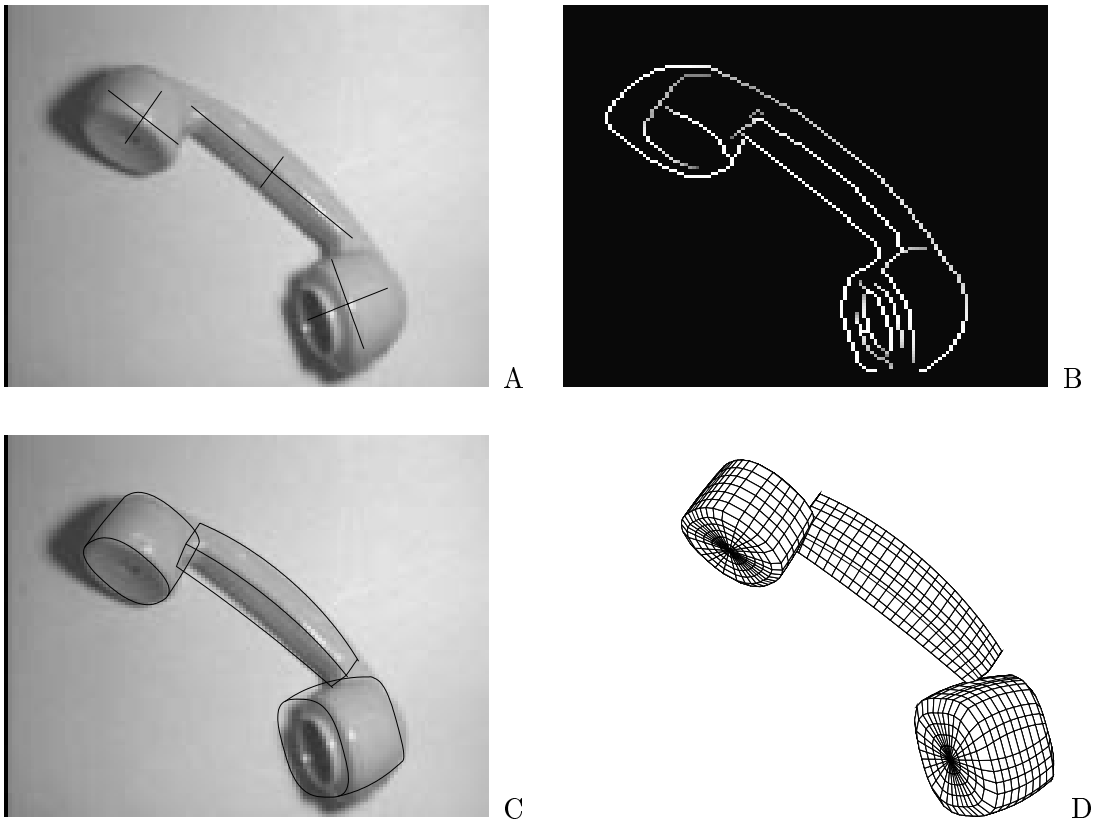
6.6.1 and fitted to each of the eight objects, with the same optimization set-up as the one given in Section 6.6.3; the resulting scores were put in a confusion table (Fig. 6.11) whose lines represent the scores of fitting an object with all the aspects. These results validate the two main assumptions of the aspect-based control strategy outlined at the beginning of Sec. 6.6: the boxed scores on the diagonal are the best ones for each geon, that is the correct aspect obtained the best score in all cases. Fig. 6.11-B show the superquadric representation using the very same parameters that result from the fitting of the best aspect. It is worth pointing out that the superquadrics are built using the very same parameters produced by the fitting and used for contructing the PDCM; these volumetric representations are then the ones that once projected onto the image plane would yield the fitted object contours.

An interesting behavior also crops up from the analysis of the scores in the confusion matrix. Let us take the case of Obj#8. The second best score corresponds to the one obtained with Aspect#6, which has a visible bottom end, whereas the third best score correspond to Aspect#4, which is the one that presents a visible bottom face but non-squared cross-section; evidently these features matched well the image and contributed to improve the overall score. Similar considerations can be made for other objects. This behavior suggests a side-effect of this strategy, that is the ranking of aspect hypotheses according to "how well" they fit the instance of objects, at least insofar as synthetic images go. With real images this phenomenon is much smoothed but is still present, as we shall see for the other examples.

6.7.3 Real image: a handset

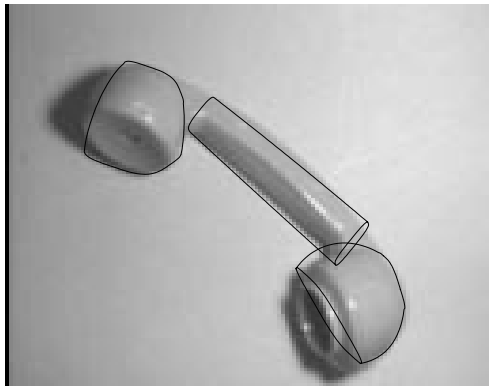
In this experiment, the now familiar 128x128 grey-level image of the handset is used (Fig. 6.12-A). The corresponding edge image is reported here for convenience in Figure 6.12-B. The initializations are performed as outlined in Section 6.6.1 and come from selected hypotheses produced by the part-based grouping and filtering method presented in the previous chapters.

Both end-pieces of the handset have almost no eccentricity and therefore it was not possible to determine their natural major axis, which is an essential requirement of geon representation. Which of the two axes was the major one was imposed by hand, but a



	Asp. #1	Asp. #2	Asp. #3	Asp. #4	Asp. #5	Asp. #6	Asp. #7	Asp. #8
Upper	-216.78	-193.43	-150.51	**-246.30**	-171.89	-88.31	-158.49	-112.34
Mid	-226.07	-223.15	-223.18	-228.82	**-386.32**	-306.45	-382.91	-301.60
Lower	-213.79	-238.68	-74.19	**-286.10**	-171.95	-156.41	-160.54	-246.30

Figure 6.12: Real-image experiment with the aspect-based control strategy. Here, the PDAs have been initialized automatically from some of the hypotheses produced by the part-based grouping and filtering method presented in previous chapters. The figure shows initialization (A), edge image (B), contour fits (C) and their volumetric representation (D). The scores of the PDA fittings are shown in the table. See text for more details.


A

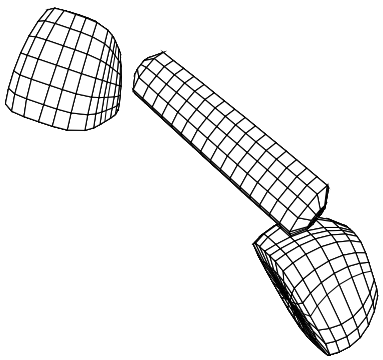

B

Figure 6.13: Handset fitting results *without* using the aspect-based strategy and from an initialization where pan, tilt and squareness values are set to 0.0, 0.0 and 0.5, respectively, and size/position/orientation as the ones in Figure 6.12. The fitting in all three cases got stuck in deep local minima.

straightforward automated strategy would just assume that for low-eccentricity blobs both major-axis hypotheses should be tried out and the best be selected. Such problems are common when the data has rotational symmetries but the models employed are oriented [Leou & Tsai 87].

As in the experiments with the synthetic image, Aspect#1 through Aspect#8 were fitted to the image for each of the initialization hypotheses, again with the same optimization set-up as is Sec. 6.6; the scores for each fit are given in the table in Fig. 6.12. The best fits, which correspond to the boxed scores, are displayed both as contours overlapped onto the real image and as superquadrics in 6.12-C and Fig. 6.12-D, respectively.

The two correct aspects for the mouth and ear pieces got the highest score as expected, since their ends are well visible. The interpretation of the mid-part has turned out to be a bit ambiguous, with the two scores for Aspect#5 and Aspect#7 very close; this is due to the invisibility of its ends and their overall low weight for such an elongated part. The correct aspect scored the highest here but either would have acceptable, given that in this case they are almost indistinguishable. It is worth remembering that we are looking for qualitative features of parts and what really matters is that the model with the right features is selected over other possible alternative ones, that is, the fitting quality need not be absolute but relative.

The aspect-based strategy helps avoid (but does not eliminate, as discussed see Section 6.7.4) situations as the real case presented in Fig. 6.13, where the fitting results are shown that are obtained from the same position/axes initialization as above but when all the parameters governing the aspect topology were left unconstrained (of course always within meaningful ranges). In the experiments of Section 6.7.1, the fitting was performed by giving a good initialization and good results were obtained; here, pan, tilt and squareness values are set to 0.0 and 0.5, respectively.
Although the number of iterations was increased to compensate for the bigger search-space, the results obtained are rather poor. The top piece is completely misinterpreted, as well as the mid-part, which was recognized as a cylinder. It can be noticed that in these two cases the fitted models (Fig. 6.13-left) match very well a considerable amount of the edges, a clear indication that very deep minima of the objective function were found there which the optimizaton algorithm could not escape. The use of topologically distinct aspects has not only the property of reducing the dimension of the search space but also of dramatically boiling down the presence and effect of undesirable minima within it.

6.7.4 Failures

As every vision researcher knows, fitting complicated deformable models by non-linear optimisation is a tricky problem indeed, and probably it will not be solved for many years to come.

In related literature, there is a chronic absence of thorough robustness tests with respect to the initial model parameters or the optimiser set-up; furthermore, only few (normally two/three) test objects are used for the experiments. Probably, we are at such an infant stage of the research that most researchers feel that it is enough to suggest new solutions, hoping that in the next few years some breakthrough will revive and select the best amongst them, as it often happens in technological evolution.

Clearly, the method for fitting deformable models proposed in this chapter can be seen as yet another of them. However, in it there are original ideas such as *i)* dropping the clumsy deformable superquadric model, *ii)* the deformable aspects with closed-form topology control, *iii)* the information-theoretical error-of-fit function, and *iv)* the

Test part	#Runs	#GOOD	#POOR	#FAILURE
Handset/Top Piece	30	15	7	8
Handset/Handle	30	20	4	6
Handset/Bottom Piece	30	11	10	9
Figure 6.9/Obj #1	20	10	7	3
Figure 6.9/Obj #2	20	11	7	2
Figure 6.9/Obj #3	20	12	6	2
Figure 6.9/Obj #4	20	3	5	12
Figure 6.9/Obj #5	20	8	10	2
Figure 6.9/Obj #6	20	4	6	10

Table 6.3: PDA fitting results for several runs with random perturbations added to an initial position. It can be noticed that currently the failure rate is rather high. See text for discussions.

embedding of simple perceptual criteria in the model prior probability. Despite that, the method is, like any other, far from being robust and usable with confidence.

Table 6.3 shows the results of the "blind" experiment that I have carried out to assess robustness to initialisation. Starting from an initial position and with the correct aspect topology, random additive perturbations were applied to the size, position and orientation parameters of the PDA, followed by the usual fitting to the image; the perturbations were Gaussian distributed with a variance equal to 10% of a_x, a_y and a_z, $N/30$ pixels for p_x and p_y (N is the resolution of the image), and $\pi/30$ for θ_{opt}; note that the noise variance for each of the parameters is several times smaller than the search bounds as given in Section 6.6.3 and therefore the optimizer should be able to yield the correct fitting starting from any of the randomly produced initializations. A large number of fitting runs were automatically executed and the results stored for each part of the handset and the six test objects in Fig. 6.9; later the fitted PDCM were visually categorised in *1)* `GOOD` when a good fitting was achieved; *2)* `POOR` to indicate that the final result was not significantly better than the *initial* position; and *3)* `FAILURE` when the final fit was totally wrong.

It can be seen that the robustness is not at all exciting. The main problem lies in the optimisation algorithm that gets stuck in deep minima; Adaptive Simulated Annealing is a very efficient stochastic optimiser and therefore it is reasonable to say that other less

sophisticated methods such as the Levenberg-Marquandt with random perturbation (such as used in in [Lowe 91] and [Borges 96]) would certainly perform no better.

Currently, however, these results cannot be compared to other methods, including works on fitting superquadrics such as [Borges 96], [Solina & Bajcsy 90], [Wu & Levine 94], [Raja & Jain 92b], [Pentland 90] and [Metaxas *et al.* 93], because no kind of robustness information was made available in these publications. Moreover, as will be re-iterated in the next section, no work was previously done in fitting deformable models of the kind of the PDA proposed here to 2D *unsegmented* edge images.

6.8 Discussion

In this chapter a novel approach to 3D qualitative part recovery from real 2D images has been presented. A new efficient deformable model is fitted to raw edge images in the framework of Model-Based Optimisation, with an objective function expressed in Bayesian terms, and the use of topologically distinct aspects has led to more reliability. The results presented here show that this method is potentially valid and open to further developments.

In this section, the major contributions of the material presented in this chapter are highlighted, followed by some criticisms and the proposition of future work.

6.8.1 Contributions

There are several contributions to vision research in this chapter. All of them were recognized by anonymous reviewers of a paper based on the chapter and its early version that appears in [Pilu & Fisher 96d].

- A new approximate but efficient parametric model of deformable superquadric contour is presented in Section 6.3. Previously, when performing deformable superquadrics fitting to 2D images as in [Metaxas *et al.* 93], a very clumsy method was used whereby the whole superquadric was built, deformed and its contour computed by finding zero-crossings of the surface normal component along the

optical axis. Here, a leaner, simple geometrical model has been pragmatically designed that approximates the contour of the deformable superquadric in a tiny fraction (about one hundredth) of the cost needed by the other method. The parameters keep a clear three-dimensional interpretation, as well as for deformable superquadrics.

- The fitting of the aforementioned parametrically deformable contour model is performed through model-based optimization where an objective function is minimized that, in information theoretical terms, essentially expresses the economy obtained by representing groups of edgels in the image by the model contour. Although similar cost functions had been proposed in the past, the one presented here formally accounts for both matched and unmatched contour portions and the background in formal Bayesian terms, whereas previous methods (such as [Fua & Hanson 89, Fua 89]) did not do so. Through experimental results, it has been shown that this method copes with a significant amount of cluttering, although there are several problems concerning robust optimization (Sec. 6.7.4).

- Although its contribution has not yet been well quantified, the embedding into the model prior probability of a bias towards more perceptually plausible 3D shapes – described in Sec. 6.5.2 – is a rather clever idea, as remarked by F. Ferrie in a personal communication.

- The concept of using an aspects-based strategy to deformable contours model fitting has been introduced here for the first time. Previous work had used aspects only for fitting CAD-based models, such as in [Eggert *et al.* 95]. The benefits of such a strategy are straightforward: the optimization can independently focus on regions of the parameter space that correspond to models with the same topology, thereby reducing the chances of getting stuck in local minima caused by different interpretations of image features. Due to the simplicity of the geon model defined in this chapter, a closed-form solution for the aspect cell subdivision has been found.

- The idea of recognizing generic primitives like geons from 2D images by fitting contour of superquadrics is not a new idea, but the only implementation known to the author is by [Metaxas *et al.* 93]. However, there the fitting was performed

to segmented data and optimization was done in image space (see Sec. 6.2.1) in a multistage fashion with two *ad hoc* search strategies for cylindroids and prismoids – probably due to severe fitting problems, also highlighted by the apparent syntheticity of the examples shown in their paper. Here, this topological information has been brought to the fore by employing distinct models, which has allowed us to safely utilize a more general optimization algorithm such as Simulated Annealing.

6.8.2 Criticisms of the method

There are several issues that need to be addressed in future work to improve the proposed method. Part of the discussion is based on comments made in anonymous reviews of a paper based on this chapter; the overall opinion on the work was rather positive, but some acute criticisms were pointed out.

The first criticism was that the method does not constitute a significant advancement with respect to the current state-of-the-art work by [Metaxas *et al.* 93]. This criticism was mainly attributed to the manual initialisation phase that was then used – now taken over by the automatic part-based grouping presented in the previous chapters. In my opinion the criticism in unjustified. The method proposed in [Metaxas *et al.* 93] assumes that faces and edges belonging to a single part are pre-segmented by the OPTICA [Dickinson *et al.* 92b] system, which also supplies information about the class of object to be fitted; this allowed them to implement an *ad hoc* strategy for dealing with the fitting of different classes of models. The problem of fitting to unsegmented data was not even taken into consideration, whereas here the fitting is done to *unsegmented* data and, in principle, the initialisation could come from methods other that the one proposed in previous chapter. Another important remark is that in [Metaxas *et al.* 93] a clumsy method for determining superquadric contours was used, whereas here a purposely designed model (Sec. 6.3) has been built that allows much greater efficiency.

Differently from the OPTICA system equipped with the superquadric fitting machinery of [Metaxas *et al.* 93], the scope of this thesis was not to build a generic-part segmentation recognition system – quite beyond the state of current vision technology – but to explore the possibility of using a global model-guided method to segment out

generic parts from ordinary edge image. The imposition of structure on the solution by the parametrically deformable aspects of the geon fitting method presented in this chapter is nothing but a natural extension to the *fil rouge* of the previous chapters.

Another criticism coming from another anonymous reviewer was that the quality of the fitting results was *not* impressive; this is a rather unfair statement that probably was inspired by inappropriate comparisons between the results given here and parallel works on part segmentation from (often pre-segmented) **range data**, such as [Solina & Bajcsy 90, Wu & Levine 94]. The absence of precise models, image cluttering and, again, the use of unsegmented 2D edge data, would never allow a precise fitting, unless other information is used.

The use of a neighbourhood "in/out" criterion in the design of the cost function of Section 6.5 has allowed a formal expression in Bayesian terms of the goodness of fit but some troubles can be encountered when the geon being fitted cannot be properly represented by the PDCM given in Section 6.3. This representation problem is common to all global deformable models fitting methods but it manifests itself more when censored error norms are employed, like [Fua & Hanson 89] [Darrell & Pentland 95] [Leonardis *et al.* 95] or those presented in this chapter.
The possibility of using a different, smoother error norm that would avoid these problems is under investigation. Preliminary experiments showed that the results are much worse than the one presented in this chapter but it too early to draw conclusions.

Some doubts could arise regarding the model prior probability given in Sec. 6.5.2. The definition might look arbitrary but it should be remembered that it was meant to have an heuristic character. In early stages of the work, this probability was uniform over the parameter space, as is often done in these cases. When such a heuristic was added, the fitting results improved rather significantly especially in regard to the recovery of 3D shape and not only the matching of the contour; further systematic experiments are needed to evaluate how these probabilities affect the final results but, given the stochastic nature of the optimiser (simulated annealing) a large amount of experiments are needed to objectively evaluate the effectiveness of such a heuristic. However, this activity was not deemed relevant at this stage and is left to future work.

Finally, and most importantly, Section 6.7.4 showed that the robustness of the method

to initialisation variation is not yet satisfying. Many factors contribute to this poor performance, the most important being the difficulties faced in optimisation stage. In sincerity, I do not feel like conjecturing about a possible solution to this: strongly-non-linear optimisation in parameter space will probably always suffer from these convergence problems. All this makes me wonder whether the model-based optimisation path followed in this chapter is the correct one and even if deformable models will ever be usable for high-level vision tasks such as part-based recognition, as we know that the ultimate aim is to achieve robustness and speed.

6.8.3 Future work

The technique presented in this chapter has opened some problems and interesting perspectives alike. Some issues that would need to be addressed in the near future are the following.

First above all, in previous chapters codons were presented as indivisible pieces of information and in Section 5.5.2 some of the limitations of such a choice were discussed. Here we are working on raw data, because after the hypotheses generation phase we have only a rough idea about which codons make up the actual part outline, let alone interior edges. In Section 6.5.1 we saw that edgels falling within a certain neighbourhood were considered as matching the model but some effort could perhaps be spent in trying to use whole codons as data to be matched. This modification would probably prevent spurious chunks of data locally matching the model contour to fool the goodness of fit evaluation, and would also yield smoother objective functions, thereby easing optimisation. Some preliminary results in this direction look promising.

A natural extension, which would however present several theoretical problems, would be to integrate other non-edge information in the fitting, specifically in the cost function, such as coarse depth and surface orientation information as it could be produced by a shape-from-shading method [Horn 89], or even by augmenting the models by some sort of appearance-based (intensity) information, in the spirit of [Murase & Nayar 92].

Another exciting step to try is to account for interactions between parts. In Chapter 5 we saw that by taking account of many competing interpretations of local evidence, it

is possible to produce a minimal, hopefully correct, interpretation of the image. The same considerations could be done here. In the case of the handset test image, for instance, the fitting could be performed concurrently for the three parts and penalty terms could be introduced for overlapping as in the support competition method of Chapter 5. However, differently from the grouping hypotheses of Chapter 5, here the fitting and hypotheses competition would be performed at the same time and the workload would be huge.

I wish to conclude this chapter with perhaps the most relevant suggestion for future work. As said back in Section 6.2.1, the fitting is performed in parameter space and this has shown considerable fragility. However, a very exciting prospect would be to use the point-to-point correspondence method by single value decomposition by [Scott & Longuet-Higgins 91] for fitting each aspect in image space, analogously to PDM fitting of the part-based grouping phase of Chapter 4. For doing so, each PDA would need to be redesigned as a point distribution model, as done in Section 3.4 for building the generic-part PDM from superellipses. This technique might allow greater robustness, speed of convergence and tolerance to bad initialisation, due to the power of the SVD correspondence method that would *globally* find the best matches between PDA landmarks of the aspects and the data, however cluttered it might be.

Chapter 7

Conclusions

This thesis has addressed the problem of generic part-based grouping and recognition from single two-dimensional edge images following a strategy that employs generic part models at all stages.

Grouping is often intended as a general-purpose early vision stage which gathers together image features of perceptual salience, usually having a well-definable structure. The key idea behind this work is performing a purposive grouping of simple parts and these parts can be conveniently represented by generic part models. In this context, part models are used to drive computational processes in gathering global relevant information.

The purpose of the research was to investigate several issues in the proposed model-guided framework and not to build a part-based vision system. The lack of an integrated implementation is not a blemish, for the endeavour of implementing such a system would have been largely worthless: much more research effort is needed before an actual implementation is both appealing and feasible.

The issues that have been addressed are key problems that naturally arose in the proposed model-guided framework, and a chapter has been dedicated to each of them.

Efficient generic 2D p art models (Chapter 3): The models that have been considered are the ellipse, the deformable superellipse and a generic part Point Distribution Model. The PDM has been trained with deformable superellipses and then used in the rest of the thesis as a generic part model, whilst ellipses have

been used to initialise PDMs from incomplete part contours.

Instantiation of part hypotheses with single edge images (Chapter 4):
In order to use part models, a method for segmenting the edge data had to be devised. The choice has been to redundantly initialise and then fit PDMs on small seed groups of perceptually salient contour portions, the codons. This phase has been called part-based model-guided grouping.

Filtering of a set of hypotheses to yield part segmentation (Chapter 5):
The large set of part hypotheses has to be cut down to a few that are highly likely to correspond to actual parts. For this purpose a Minimum Description Length method has been employed that finds the best description of the edge image in term of generic part models.

Recover the qualitative 3D structure of the generic parts (Chapter 6):
Close-by parts can have a pronounced tri-dimensional structure. In order to recover it, parametrically deformable aspects have been used and fitted by a Maximum a Posteriori estimation to the raw edge image; the initialisations are provided by the filtered hypotheses produced by the grouping and filtering stages.

As always happens, during the investigation of these matters several spin-off issues were identified and successfully investigated, such as a major contribution to the ellipse fitting problem, a new parametric method for sampling superellipse models, the training of PDM models on other models, and an extremely efficient model for approximating the projected contour of deformable superquadrics.

In the following, I summarise the major contributions, some criticisms, and future work, referring to the respective chapters for more detailed discussions.

7.1 Contributions

This section summarises the original contributions to computer vision made by this thesis. A heavily abridged paper covering all these issues appears in [Pilu & Fisher 96a].

Ellipse-specific direct least squares fitting. Section 3.2 presents the first direct method for specifically fitting ellipses in the least-squares sense. Previous approaches have used either generic conic fitting or relied on iterative methods to recover elliptic solutions. The proposed method is *i)* ellipse-specific, *ii)* directly solved by a generalised eigen-system, *iii)* has a desirable low-eccentricity bias, and *iv)* is extremely robust to noise.

Fitzgibbon [Fitzgibbon & Fisher 95] devised the method but was unaware of its importance. I soon spotted its originality and significance, which I confirmed by an extensive literature review, and set about providing theoretical justification for it.

This work appears in [Pilu *et al.* 96a] and in [Fitzgibbon *et al.* 96] with more quantitative experiments.

Equal-distance sampling of superellipse models. Appendix A provides a new parametric method for achieving equal-distance sampling of contours of deformable superellipses, with obvious extension to superquadrics. Superellipses are parametric models that can be used for representing two dimensional object parts or aspects of 3D parts. Previously, little care was given to obtaining a precise sampling of the contours of these models. However, equal-distance sampling of superellipse model contours is important for several reasons, and in particular, I needed it for building a better statistical model of the contour of a deformable superellipse in order to more efficiently represent it by Point Distribution Models (see below).

An extended version of this work appears in [Pilu & Fisher 95].

Training PDMs on Deformable Superellipses. Section 3.4 addresses the following problem: How can we make a complicated mathematical shape model simpler while keeping a comparable representational power? The proposed solution is to use the original model itself – which represents a class of shapes – to train a Point Distribution Model (PDM). In the context of this thesis, the method is applied to the case of deformable superellipses, which are suitable models for coarsely approximating generic part outlines.

This part appears in [Pilu *et al.* 96b].

Part-based model-guided grouping. Within our model-guided strategy, our generic (deformable) models are used to represent part hypotheses and need to be fitted to unsegmented edge data. All previous fitting methods using deformable models performed fitting after manual initialisation, except in the simplest cases. Here, thanks to the (intentionally) simple and *self-symmetric* (see Sec. 3.4.5) nature of our part models, automatic instantiation of a redundant set of model hypotheses has been possible and a general method is presented in Chapter 4. Briefly, small sets (*seeds*) of codons are used to first *pre-shape* the deformable models, followed by a full growing stage using additional evidence found in the image. The fundamental assumption is that for our simple models, a few, well chosen features give enough information to recover coarse part structure. Remarkably, such an approach, a familiar one in the traditional vision literature, has (to my best knowledge) never been proposed for deformable models. Currently, pairs of codons are used for pre-shaping models, leading to $O(N^2)$ complexity in the number of codons N. This stage does not produce a definitive part segmentation (except in the simplest cases) but a set of part hypotheses. An abridged description of this grouping strategy appears in [Pilu & Fisher 96a].

Part filtering by the MDL criterion. Chapter 5 presents a novel method for filtering the redundant set of 2D part hypotheses (those produced by the previous grouping stage) that retains only those that are likely to correspond to actual parts. The method is inspired by recent work in segmentation using the Minimum Description Length (MDL) criterion, which has previously been used for segmenting surfaces into patches. For the first time here, the philosophy is applied to a two-dimensional context. In the proposed method, supporting evidence for hypotheses is put into competition under the MDL framework to select part hypotheses that most economically represent supporting edges in the "language" of generic parts. The filtering is performed by the maximisation of a quadratic boolean cost function by a genetic algorithm. Numerous experiments are provided that show the stability of the method but also stress a few inherent limitations, summarised in Section 7.2. Notably, the psychological experiments presented in Appendix E shows that the part segmentation thus obtained is comparable to that produced by human subject when not told what is the part model

to be used.

An shortened version of this part appears in [Pilu & Fisher 96b] and has been submitted for journal publication as [Pilu & Fisher 96c].

Recovery of 3D structure by parametrically deformable aspects. Chapter 6 presents a novel approach to the recovery of the structure of generic solid geon-like parts of objects from real 2D images. Most previous work on detection and recognition of geons from 2D images has relied on quasi-perfect line drawings. The use of aspects has also been proposed for CAD-like models. Here, I introduced the concept of *parametrically deformable aspects* (PDAs) as 2D models to be matched to real images of geons in the framework of Model-Based Optimisation. Parametric models efficiently represent geons, and the use of topologically different aspects yields a more robust fitting, which is performed by a maximum a posteriori (MAP) estimation and a simple aspect-based control strategy. The MAP estimation is defined on sound theoretical grounds and a model prior probability is introduced that biases fitting towards perceptually plausible models. The experiments of Section 6.7.4 point out problems of robustness to initialisation variation which have not been analysed in similar papers. The proposed method is general, in the sense that it could be easily applicable to other parametrically defined part vocabularies.

An earlier version of this chapter that did not use the aspect-based control strategy appears in [Pilu & Fisher 96d]; the whole chapter is available also in [Pilu & Fisher 96e]. In both publications, however, no robustness experiment is included.

7.2 Criticisms

Sections on criticisms and limitations of each stage have been put at the end of each chapter. In order to avoid verbosity and repetitions, here I cite just the major issues. I shall use a brief question-reply format where I put forth criticisms and succinctly give replies. Many of these matters were raised in casual discussions with colleagues.

Modelling

Aren't your models a bit too simple? Yes, they are. Ellipses and, to a lesser extent, the proposed PDM, are very coarse representations of real parts. However, in the spirit of [Marr & Nishihara 78, Biederman 87], basic categories need not be much more complicated.

Isn't this "training PDMs on models" idea a trivial one? Yes, it could be seen as trivial. However, the motivation is the fundamental observation that often a model is chosen for certain representational characteristics regardless of fitting problems. Here, I propose to use PDMs that function as the original models themselves, but are easier to fit.

Grouping by models

Why should we use models at all for performing a qualitative task such as grouping? Because using generic models is a way of "imposing a structure" on the problem. However, models do have drawbacks in their limited generality and the need to "find" the data to be fitted to; deformable model fitting is still a tricky business for unsegmented images.

Why not use symmetry-based or convex grouping approaches? These two approaches are extremely powerful and are likely, when fully integrated in vision systems, to produce major breakthroughs. The model-guided approach presented here is complementary to those, and addresses just the limited domain of simple geon-like part grouping.

Is the proposed method robust enough? Currently there are problems in the initialisation phase and in the pre-shaping when objects are too tapered or too bent. These problems have not been fully solved yet, but more sophisticated methods could be devised.

Hypothesis filtering

There are ambiguities. Yes. As shown in Section 5.3.7, in edge-based approaches there are some unavoidable ambiguous representations due to multiple competing symmetries and figure-ground effects.

Why don't you try to get many alternative solutions? This would answer the previous issue but it is a relatively challenging task. I plan to

explore the matter further by using a multi-population GA. In a personal communication, Ales Leonardis suggested the use of tabú search methods, which can yield several alternative good solutions.

Is it realistic to rely just on edges? No, other information should have been used. However, a coherent method for the integration of other information would be a research topic in its own right, therefore it is left for future work.

Is the approach scalable to bigger problem? At the moment I have no definite answer. Larger problems might give rise to many more ambiguities, which would be difficult to discriminate between. Despite this, it must be said that GAs are quite powerful when dealing with large boolean problems, so at least there should not be problems in the optimisation phase.

PDA fitting

How about Metaxas' work? Since Metaxas [Metaxas *et al.* 93] successfully addressed the same problem of recovering 3D shape of simple parts, my work on this problem could be seen as superfluous. However, a closer look reveals that Metaxas uses segmented images (the output of Dickinson's OPTICA system) and this method could not be possibly used in our context, where we have just coarse part initialisations and unsegmented images. Although significant, this problem (to my best knowledge) has never been addressed before.

The error of fit function (Sec. 6.5) uses too a simple distance norm. This is true. Although the use of an in/out distance norm in Sec. 6.5 has been successful in allowing a formal MAP formulation of the fitting problem, it is limited in that if the data cannot be well-described by the model, it might fail.

Did you change fitting strategy? Yes. In the grouping phase, PDMs are fitted in image space, whereas in this thesis the PDAs are fitted in parameter space, which I previously mentioned was problematic. This change reflects the timing in the investigation of the two issues, since the PDA fitting was investigated in a much earlier stage of the research. Suggested modifications are detailed in the section on future work.

The method is rather fragile. True. Section 6.7.4 showed that the robustness of the method to initialisation variations is not yet satisfying. I do not feel like conjecturing about a possible solution to this: strongly non-linear optimisation in parameter space will probably always suffer from convergence problems. By any rate, however, the PDA idea is highly original so I preferred to assess its validity in a larger context by presenting it at a conference [Pilu & Fisher 96d] before carrying out more work; feedback received from F. Ferrie, F. Solina and A. Leonardis has been very encouraging. Rather, the question that should be asked is another one: will deformable models and model-based optimisation ever be usable for high-level vision tasks such as part-based recognition, since we know that after all the ultimate aim is to achieve robustness and speed?

7.3 Further work

Interesting new issues have cropped up during the course of the research described in this thesis. Most of them are spin-off directions which could not be investigated for reason of time. Here I summarises the most significant ones. More details can be found in the respective sections or chapters.

Ellipse fitting: I plan a further analysis of the new ellipse-specific direct least squares algorithm in order to *i)* theoretically characterise its noise performance by using the eigenvalue perturbation theorem [Wilkinson 65]; *ii)* assess its benefits when used as a generator of initial estimates for iterative fitting methods; and *iii)* develop a bias correction method (perhaps following [Kanatani 94]).

Training models on models: The method can be extended (if needed) to deformable superquadric models, also with more domain-specific or local deformations. A suggestive idea, which is certainly worth investigating for its own sake, is to "train" a parametric representation of a high order polynomial, which can be fitted directly by least squares methods.

PDM fitting: I caught a glimpse of an interesting possibility, which is fitting PDMs in the least squares sense to scattered data points in a much more efficient and

robust way by integrating Eqn. (3.13) with the feature correspondence method of [Scott & Longuet-Higgins 91]. By doing so, the troublesome PDM initialisation phase (currently done by ellipse fitting) could perhaps be dropped.

Codon pre-grouping: Currently, the pre-grouping phase consists of exhaustively generating pairs of codons. This phase is crucial for the generation of the right part hypotheses. The use of heuristics based on perceptual organisation (e.g. symmetry and convex grouping) would dramatically increase performance, allowing to cope with considerably bent and tapered shapes, and (most probably) reduce complexity.

Filtering by MDL: The integration of other information is needed if the method is to become really robust. In the simple edge-based approach, more work could be done to better deal with ambiguities, perhaps by eliciting them as part of the nature of the problem, rather than just an annoyance. A clever engineering of the objective function evaluation and optimiser (GA) could increase time performance several-fold.

PDA fitting: If needed, further work could improve this stage dramatically. In particular, following the strategy of training PDM on models, each PDA could be modelled by a properly trained PDM. By doing so, the fitting could be performed in image-space and therefore likely to become faster and more robust, especially in conjunction with the feature correspondence that could be achieved by [Scott & Longuet-Higgins 91], which would also facilitate dealing with interior model edges. Finally, the PDA idea could be extended to more complicated deformable models.

Appendix A

Equal-Distance Sampling of Superellipses

This appendix[1] presents a study on the problem of sampling at equal distance the contour of superellipses.

Superellipses and superquadrics are mathematical objects of a particularly awkward nature because of strong non-linearities caused by fractional exponents, which cannot be easily dealt with analytically .

In many works, superquadrics were not sampled in a regular way when rendered or when an objective function is to be computed across superquadric surfaces (e.g. [Wu & Levine 94][Borges 96]); in the latter case, irregular sampling causes some regions to have a higher weight on the final cost, contributing to wrong results with real data.

In [Franklin & Barr 81], this problem was partially solved by using an explicit non-parametric method which greatly improves precision but still gives 20-30% error in the sampling distance and cannot deal with deformations. In that work, parametric sampling was ruled out because of its complexity and slowness.

Here, I present a parametric technique for equal-distance sampling the contour of superellipses which is fast and reasonably accurate; an extension to deformed models is also suggested. By using the spherical product [Barr 81] , the method can be trivially extended to superquadrics.

A.1 Linear sampling and Franklin's explicit method

This section illustrates two contour sampling methods, namely *linear sampling*, which uses a plain linear increase of the θ parameter in Eqn. 3.8, and the explicit method proposed by in [Franklin & Barr 81].

The top of Figure A.1 shows a linear sampling of the superellipse parameter θ (left)

[1] An extended version of this appendix appears in [Pilu & Fisher 95].

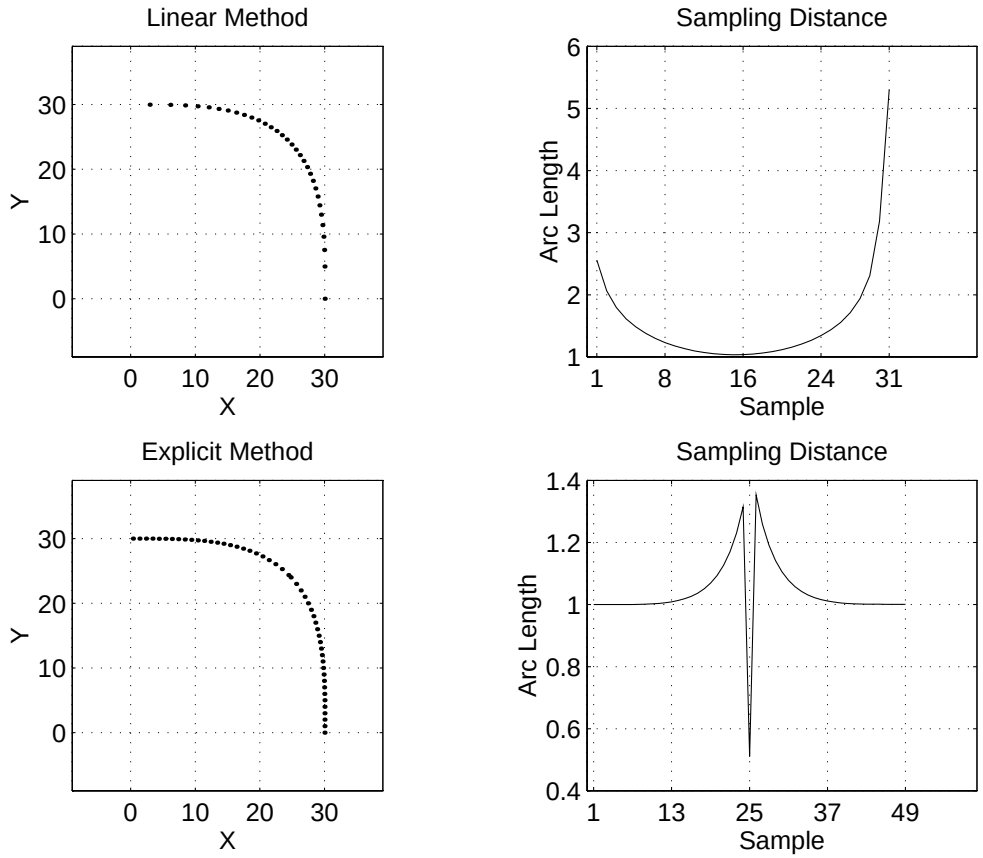


Figure A.1: Example of linear parameter sampling (top) and Franklin's explicit sampling (bottom); the two left figures show the sampled points in the first superellipse quadrant, with the respective sampling distances given on the right-hand side. Although the explicit method fares better, it still gives high sampling distance variations.

and a graph expressing the distance of successive samples (right). It can be easily seen that this method, although fast and simple, has an important limitation: points are very evenly spaced and mostly gathered near the corners, where large changes of θ involve small changes in both x and y. For higher values of the squareness parameter ϵ, this phenomenon is even more accentuated.

The bottom of Figure A.1 shows a sampling example by the method proposed in [Franklin & Barr 81], where one of the coordinates is assigned and the other is computed by solving Eqn. (3.9) for the former. This sampling technique is considerably better (see Fig. A.1-bottom-right) than the plain linear method but, as it can be seen from the graph of the distance (Fig. A.1-bottom-left), it still gives more than 30% error. (The spike is due to a mismatch between the two halves of the quadrant at the junction.) Moreover, this method relies on a straight approximation of the two halves of the superellipse quadrant and therefore cannot deal with any kind of deformation, which would only worsen the distance spread along the contour.

A.2 Optimal parametric sampling

In order to avoid the high discrepancy of sampling distance on the superellipse contour, a simple first-order differential model can be employed to control the sampling according to local contour curvature.

First order model (Model 1)

Consider the parametric equation of a superellipse as in Eqn. (3.8) and let $\mathbf{x}(\theta) = [x(\theta)\ y(\theta)]^T$ be a point on the superellipse contour. We can approximate the arclength between two close points $\mathbf{x}(\theta)$ and $\mathbf{x}(\theta+\Delta_\theta(\theta))$ by the length segment linking them:

$$D(\theta)^2 = |\mathbf{x}(\theta + \Delta_\theta(\theta)) - \mathbf{x}(\theta)|^2$$

Assuming a relatively small $\Delta_\theta(\theta)$, the right hand side of this equation can be approximated to first order by:

$$D(\theta)^2 = \left(\frac{\partial}{\partial\theta}(a_1 cos(\theta)^\epsilon)\Delta_\theta(\theta)\right)^2 + \left(\frac{\partial}{\partial\theta}(a_2 sin(\theta)^\epsilon)\Delta_\theta(\theta)\right)^2$$

By expanding and solving for Δ_θ we obtain:

$$\Delta_\theta(\theta) = \frac{D(\theta)}{\epsilon}\sqrt{\frac{cos(\theta)^2 sin(\theta)^2}{a_1^2(cos(\theta)^\epsilon)^2 sin(\theta)^4 + a_2^2(sin(\theta)^\epsilon)^2 cos(\theta)^4}} \tag{A.1}$$

If an equal distance sampling is desired, $D(\theta)$ must be set to a constant $\overline{D}$ that approximately represents the desired arclength between two sampled points; $D(\theta)$ could also been adaptively changed for different kind of samplings or to cope with deformations.

The two dual updating algorithms for θ (in the first quadrant) should then be as simple as:

$$\theta_i = \theta_{i-1} + \Delta_\theta(\theta_i) \qquad \theta_0 = 0\ , \quad i \in \{1..N\} \text{ such that } \theta_N < \pi/2 \tag{A.2}$$

$$\theta_i = \theta_{i-1} - \Delta_\theta(\theta_i) \qquad \theta_0 = \pi/2\ , \quad i \in \{1..N\} \text{ such that } \theta_N > 0\ , \tag{A.3}$$

the former going up step by step from 0 to $\pi/2$ and the latter from $\pi/2$ down to 0 for the first quadrant, where N is the number of samples per quadrant; other quadrants can be trivially obtained by mirroring.

Unfortunately, the strong non-linearities of superellipses cause this approximation to be wrong for θ close to 0 and $\pi/2$, and even the apparent equivalence of the sampling schemes (A.2) and (A.3) breaks down. In fact, the sampling distance increases as θ increases due to the first order (linear) approximation that has been employed: in regions of *increasing* curvature, the computed derivative overestimates the rate of change

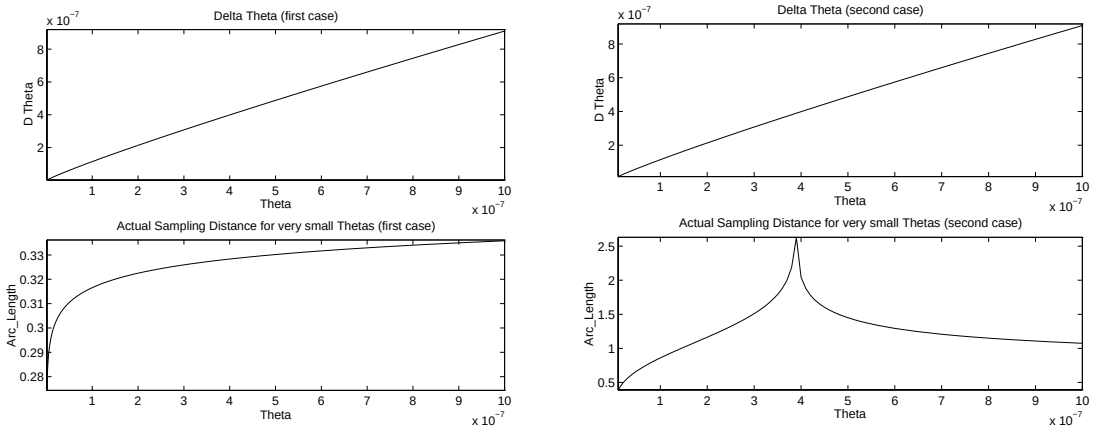


Figure A.2: Actual sampling distance for small θ. See text for details.

of θ needed to obtain a certain arclength, whereas in regions of *decreasing* curvature the reverse happens. As a result, the real arclength is much lower that it should be in some regions and much higher in others. Figure A.2 highlights this effect for very small θ (in which the rate of change in the curvature tends to infinity) in the case of sampling scheme (A.2) ("first case", on the left) and (A.3) (on the right). It is noticed that the second case ("second case", right) is equivalent to sampling with scheme (A.2) used near $\pi/2$. In the first case θ goes to zero very quickly whereas in the second it tends to infinity but once $\theta - \Delta_\theta$ goes below zero, its behaviour inverts and becomes similar to the one in the first case.

Dealing with singularities (Model 2)

In order to avoid problems at singularities, it has been found that the following simple model yields a very good approximation to the equal-distance sampling near the singularities $\theta = 0$ and $\theta = \pi/2$.

In the case with $\theta \to 0$, Eqn. (A.1) can be approximated as:

$$\mathbf{x}(\theta) = \begin{bmatrix} a_1 \\ a_2\theta^\epsilon \end{bmatrix}$$

and hence the distance between two points in this case is therefore:

$$D(\theta) = y(\theta + \Delta_\theta(\theta)) - y(\theta) = a_2(\theta + \Delta_\theta(\theta))^\epsilon - a_2\theta^\epsilon$$

By solving for $\Delta_\theta(\theta)$ we obtain:

$$\Delta_\theta(\theta) = \left(\frac{D(\theta)}{a_2} - \theta^\epsilon\right)^{\frac{1}{\epsilon}} - \theta \tag{A.4}$$

Analogously, for $\theta \to \pi/2$ we have:

$$\Delta_\theta(\theta) = \left(\frac{D(\theta)}{a_1} - (\pi/2 - \theta)^\epsilon\right)^{\frac{1}{\epsilon}} - (\pi/2 - \theta) \tag{A.5}$$

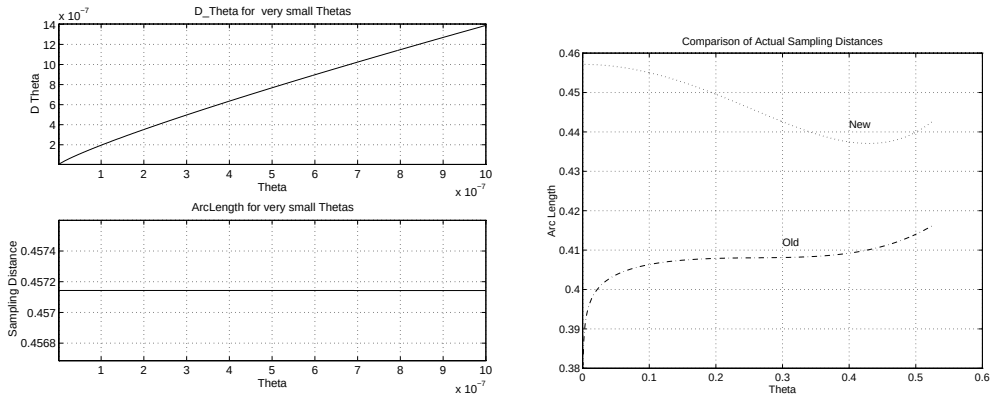


Figure A.3: New approximation for small θ (left) and a comparison to the previous method for larger θ (right). See text for details.

with $D(\theta)$ set to a constant $\overline{D}$ if an equal-distance sampling is desired.

Figure A.3-left shows how this new model behaves for very small θs; again we have $\epsilon = 0.1$, $a_1 = 20$, $a_2 = 20$. A quick comparison with Figure A.2 shows that the actual distance with this new sampling model is practically constant with θ, which is what we wanted to achieve. For larger values of θ, as expected, this approximation does not hold any longer and some small errors are introduced (A.3-right). It has to be noticed, however, that here we have used a low value of roundness ($\epsilon = 0.1$) and this is why the small-θ approximation holds even for relatively large values of θ. For rounder shape this approximation will hold for smaller and smaller values of θ but, at the same time, the distance Eqn. (A.2) will become more and more suitable because the non-linearities becomes less strong.

Combination of the two models

So, Model 1 (Eqn. (A.1)) is the generic first order model of the sampling distance and Model 2 (Eqn. (A.4)) is the one which is to be used near the singularities $\theta = 0$ and $\theta = \pi/2$.

The two models are combined by switching between the two models after θ and ($\pi/2$–θ) go below a certain threshold τ. It has been experimentally found that a good value of τ is 10^{-2}, which gives relatively smooth change in the actual sampling distance both for very small and large ϵ.

However, when (A.5) is used, a problem arises that is caused by the high numerical precision in subtraction ($\pi/2 - \theta$) needed for small values of ϵ. This problem has been bypassed by swapping the x and y axis in the superellipse Eqn. (3.8) in order to have the condition $\theta \to 0$ in place of $\theta \to \pi/2$; a_1 will be then used instead of a_2 in order to have exactly the same shape, as shown in Figure A.4-left.

Figure A.4 (right) shows the result of this sampling method for $\epsilon = 0.1$, where the small circles indicates the switching points between Model 1 and Model 2; note that

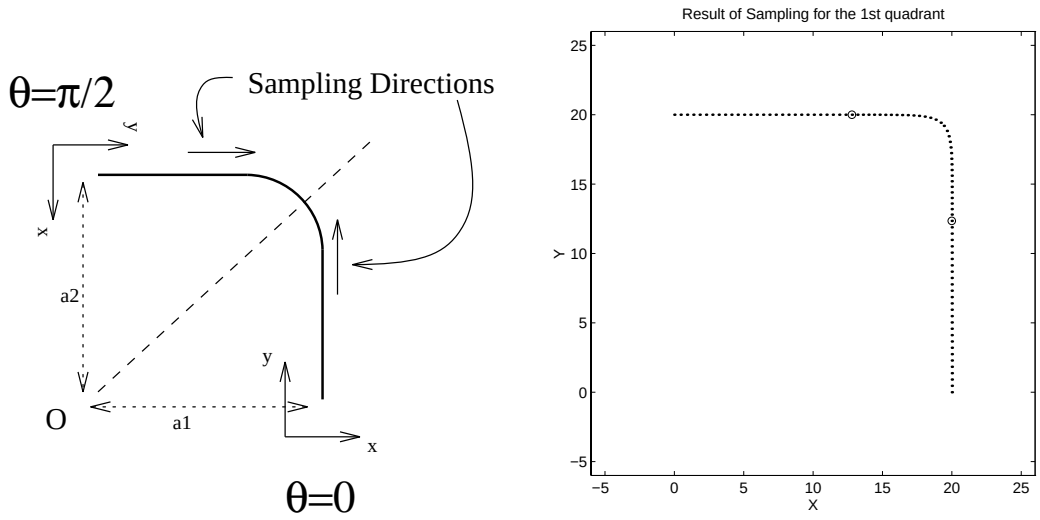


Figure A.4: Swapping of axes (left) and final sampling result.

in this example the position of these two circles represent a distance in the parameter space θ of just $\tau = 10^{-2}$ radians!)

Figure A.5 gives a full example in which θ, Δ_θ, the actual distance $D(\theta)$ and the sampled superellipse are given for $\epsilon = 0.2$, $a_1 = 20$, $a_2 = 20$. The discontinuities at **A** and **B** are due to the swap from Model 1 and Model 2; the steep spike at **C** is caused by an unavoidable mismatch of the two halves of sampling joined together as shown in Figure A.4-left. In both examples it can be seen how good the sampling is, with an error in the actual distance as low as 5% on the full $[0..\pi/2]$ range.

The full sampling of the superellipse contour for $\theta = [-\pi..\pi]$ is trivially obtained by mirroring and reversing the first quadrant.

A.3 Extension to deformed superellipses

Superellipses are of particular utility when deformed because they can represent more complex shapes. When deformations such as tapering and bending of Section 3.3 are applied, the sampling distance changes along the contour; Figure A.6 (top) shows this effect for the first quadrant of a tapered superellipse ($\epsilon = 0.5$, $K_x = 0.7$ and $\overline{D} = 3$) sampled along with the explicit method [Franklin & Barr 81] with the corresponding sampling distance. As it can be seen the error is rather big.

We show now how to extend the proposed method for tapered superellipses. (The same idea could be employed to deal with bending deformations but with more complex formulae).

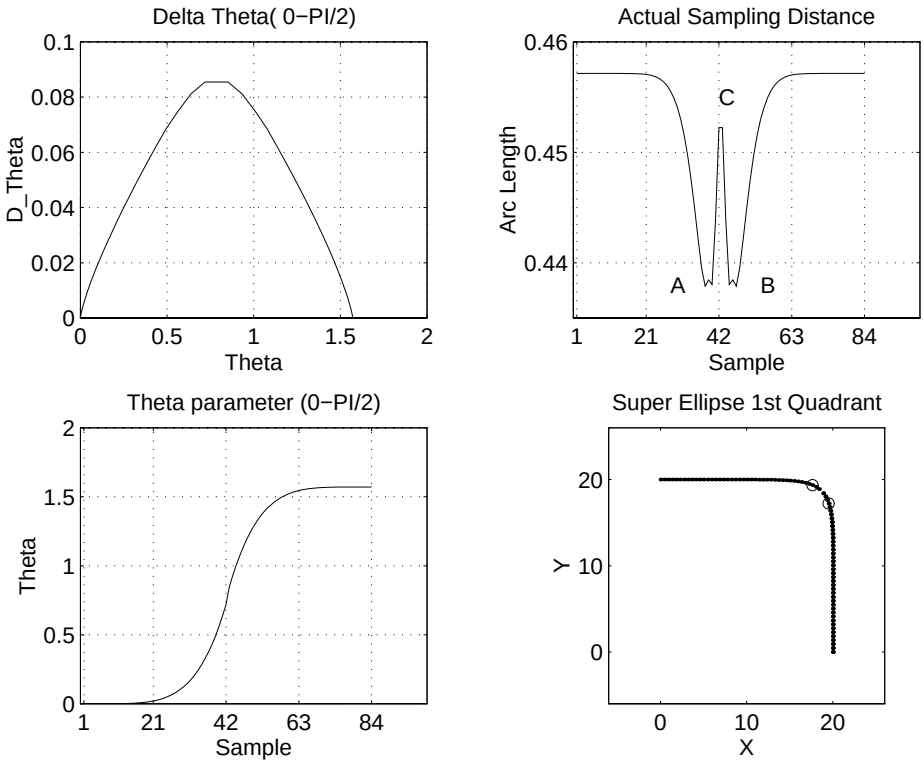


Figure A.5: An examples of equal-distance sampling (see text).

By combining Eqns. (3.8) and (3.11), the equation of a linearly tapered superellipse can be written as:

$$\begin{cases} x(\theta) = (T_x sin(\theta)^\epsilon + 1) a_1 cos(\theta)^\epsilon \\ y(\theta) = a_2 sin(\theta)^\epsilon \end{cases}$$

As in Section A.2, we can express the sampling distance $D(\theta)$ as a function of $\mathbf{\Delta}_\theta(\theta)$ and by solving for the latter we have:

$$\mathbf{\Delta}_\theta(\theta) = D(\theta)\sqrt{\left(\frac{\partial x(\theta)}{\partial \theta}\right)^2 + \left(\frac{\partial y(\theta)}{\partial \theta}\right)^2}$$

Near the singularities we need to employ a different model which is the same as Model 2 but with a modified sampled distance $\overline{D}'$ instead of $\overline{D}$ to account for what the distance is going to be *after* the deformation. By considering the tapering geometry, and assuming that near the singularities the shape is practically straight, we have:

$$\begin{cases} \theta \to 0: & \overline{D}' = \overline{D} atan\left(\frac{a_1 T_x}{a_2}\right) \\ \theta \to \pi/2: & \overline{D}' = \overline{D}\frac{1}{T_x+1} \end{cases}$$

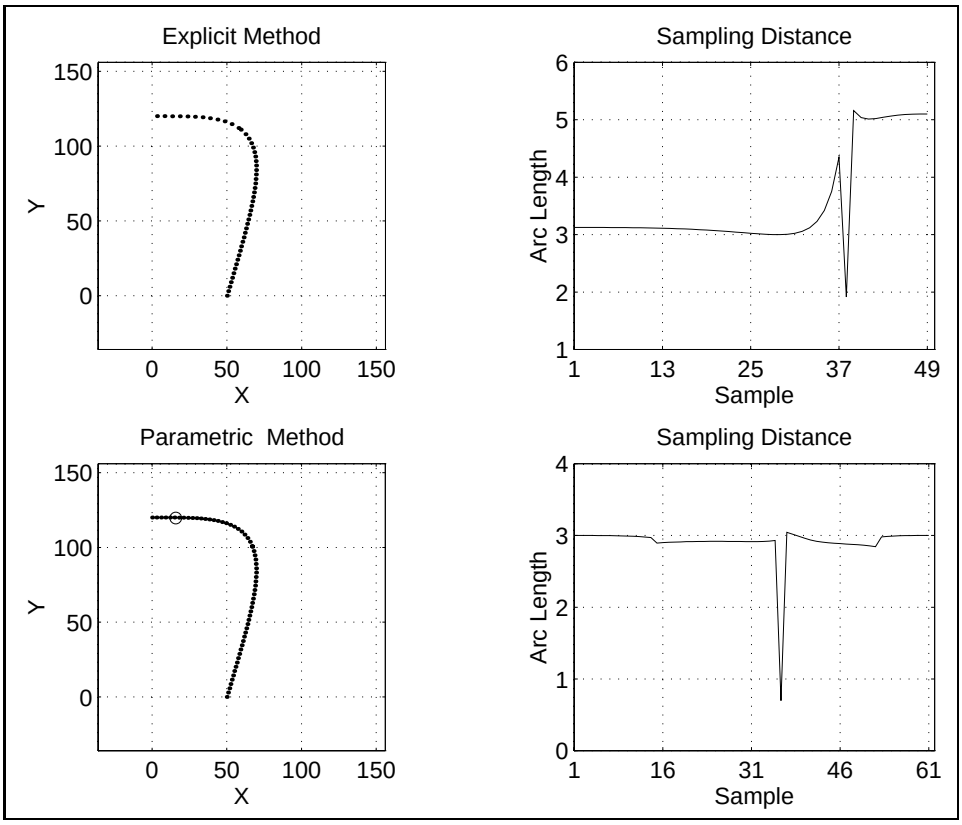


Figure A.6: Example of sampling a deformed superellipse: explicit method (top) and proposed parametric method (bottom).

Figure A.6-bottom shows the result of this sampling method in the same case as before along with the actual sampling distance; from the distance graph, it can be seen that the improvement is significant in all the $[0..\pi/2]$ range over the explicit method, shown in Figure A.6-top.

Appendix B

Feature correspondence by singular value decomposition

This appendix reviews of the algorithm for matching 2D point-features across a pair of images presented by [Scott & Longuet-Higgins 91]. Extensions to the method appears in, i.e., [Sclaroff & Pentland 95] and the elegant paper by [Shapiro & Brady 92]; in particular, what follows draws from the latter, to which I refer for further details and developments.

Following the "minimal mapping" philosophy of [Ullman 79], the algorithm incorporates the *principle of proximity* (favouring matches across shorter distances) and the *principle of exclusion* (favouring one-to-one matches). The resulting mapping is effectively minimal, in the sense that it minimises the overall sum of the squared distances travelled by the features, subject to the uniqueness constraint. A remarkable feature of the algorithm is its elegant implementation, founded on a well-conditioned eigenvector solution which involves no iterations.

The input of the algorithm is a set of m features $(x_{i,1})$ in image I_1 and (n) features $(x_{j,2})$ in image I_2. The computation then proceeds in three stages:

1. Build a *proximity matrix* $\mathbf{G}$ of the two sets of features. Each element G_{ij} records the attraction between the i^{th} feature in I_1 and the j^{th} feature in I_2 via a Gaussian-weighted distance metric

$$G_{ij} = \exp^{-d_{ij}^2/2\sigma^2}, i = 1 \ldots m, j = 1 \ldots n$$

   where $d_{ij}^2 = \|x_{i,1} - x_{j,2}\|^2$ is the squared Euclidean distance between the two features; G_{ij} ranges from 0 for widely-separated features ($d_{ij} = \infty$) to 1 for coincident features ($d_{ij} = 0$). The parameter σ controls the degree of interaction between the two sets of features: a small value of σ enforces local interactions, while a larger value permits more global interactions.

2. Find the *singular value decomposition* (SVD) of $\mathbf{G}$:

$$\mathbf{G} = \mathbf{TDU}.$$

The **T** and **U** matrices are orthogonal and the **D** matrix contains the (positive) singular values along its diagonal in descending numerical order.

3. Compute the correlation (in a scalar product sense) between **T**'s rows and **U**'s columns, giving an *association matrix* **P**,

$$\mathbf{P} = \mathbf{TEU}.$$

where **E** is obtained by replacing each diagonal element in **D** by a 1[1]. Each element P_{ij} then indicates the attraction strength between features $x_{i,1}$ and $x_{j,2}$, where 1 indicated a prefect match and 0, no match at all. The correspondence between the two features is "strong" only if P_{ij} is largest in both its row and its column, implying a *mutual consent* to the match.

In [Scott & Longuet-Higgins 91], it was shown that the above algorithm maximises the trace of $\mathbf{P}^T\mathbf{G}$. That is the **P** matrix was effectively a "mask" which slotted over **G** and selected the biggest elements. Since G_{ij} was large precisely when d_{ij}^2 was small, an overall minimum squared distance mapping was ensured.

Figure 3.16 shows two examples of the application of this algorithm to the feature correspondence of two hand-input patterns.

[1] The matrix **E** is used only for illustrative purpose since, being $\mathbf{E} = \mathbf{I}$, it is $\mathbf{P} = \mathbf{TU}$. .

Appendix C

Simulated Annealing

Simulated Annealing (SA) [Kirkpatrick *et al.* 83] is a powerful optimisation tool that effectively combines gradient descent and controlled random perturbation to perform the minimisation of non-convex functions. It was developed from the Metropolis algorithm [Metropolis *et al.* 53], which was originally contrived to simulate the evolution towards thermal equilibrium states in statistical mechanics. The Metropolis' algorithm can be summarised as follows. Given a solid composed by interacting atoms, small random perturbations are added to the current state; a differential of energy ΔE is computed and if $\Delta E < 0$ the new state is accepted as a valid one. Conversely is $\Delta E \geq 0$ the state is not rejected but it is given a probability $e^{-\Delta E/kT_a}$ (*Metropolis criterion*) , where k is the Boltzmann constant and T_a is the absolute temperature. By keep repeating this procedure for a large number of times the system eventually converges to a thermal equilibrium.

More recently, in [Kirkpatrick *et al.* 83] an important modification was proposed to the Metropolis' algorithm that consisted of running it with decreasing temperatures (called Boltzmann Annealing) until a low enough temperature is reached, analogously to the physical annealing process of a solid. The method was called *simulated annealing* and the way the temperature is lowered called *annealing schedule*.

Optimisation by SA was first introduced to the vision community in the seminal paper [Geman & Geman 84] and more recently used also in [Lim 94, Wu & Levine 94] and other works.

In this thesis, I have used a recent publicly available implementation of SA has, called Adaptive Simulated Annealing (ASA), developed by Ingber at Caltech [Les93]. As described by Ingber, *"the major difference between ASA and standard Boltzmann SA is that the ergodic sampling takes place in a $n+1$ dimensional space, in terms of n state variables and the cost function"*.

Appendix D

Polynomial Fitting

This appendix describes the least squares fitting of second order polynomial of the form $v = au^2 + bu + c$ to a list of points $\mathbf{P} = \{(u_1, v_1), \ \ldots, \ (u_n, v_n)\}$ subject to the constraint of passing through the first and last point (u_1, v_1) and (u_n, v_n), respectively.

Let us denote this fitting procedure as $\{a, b, c\} = PF(\mathbf{u}, \mathbf{v})$, where $\mathbf{u} = [u_1 \cdots u_n]^T$ and $\mathbf{v} = [v_1 \cdots v_n]^T$.

The solution can be found by writing the general equation of the polynomial passing through the two points as a function of c, and finding its least squares solution. The equations are given below:

$$
\begin{array}{lcl}
\mathbf{A} & = & \mathbf{u'}^2 u_n - \mathbf{u'}^2 u_1 + u_1^2 \mathbf{u'} - u_n^2 \mathbf{u'} + u_n^2 u_1 - u_n u_1^2 \\
\mathbf{B} & = & \mathbf{v'} u_n^2 u_1 - \mathbf{v'} u_n u_1^2 - \mathbf{u'}^2 v_n u_1 + \mathbf{u'}^2 v_1 u_n - u_n^2 \mathbf{u'} v_1 + u_1^2 \mathbf{u'} v_n \\
c & = & (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{A}^T \mathbf{B} \\
a & = & \frac{u_1 v_n + u_n c - u_1 c - u_n v_1}{u_1 u_n (u_n - u_1)} \\
b & = & \frac{+u_n^2 v_1 - u_1^2 v_n + u_1^2 c - u_n^2 c}{u_1 u_n (u_n - u_1)}
\end{array}
$$

where $\mathbf{u'} = [u_2 \cdots u_{n-1}]^T$ and $\mathbf{v'} = [v_2 \cdots v_{n-1}]^T$.

Appendix E

Part decomposition: A psychological experiment

This appendix describes the psychological experiment designed to assess both the notion we have of object parts and our ability to segment them solely from edge images.

Firstly, I outline the motivation of the experiment, followed by the description of the experimental set-up and its underpinning. Then, the experimental data collected is presented and briefly commented upon. Finally, the results are extensively discussed and related to those produced in this thesis in Chapter 5.

E.1 Motivation

In the course of this research, it was assumed that the definition of part was a clear cut thing and was generally based on the ideas sprung up in the past 25 years in landmark works such as [Binford 71], [Marr & Nishihara 78], [Hoffman & Richards 85], [Pentland 86] and [Biederman 87]. Most of these works employ a *sausage-like* model of parts to speculate upon the ways computers should recover and deal with them.

Despite this well-established belief of what a part should be, during the final stage of this thesis I was posed an interesting question: *Are your results comparable to those produced by humans asked to segment objects into their composing parts? And do different people give same judgements?*

My answer to this question was that probably the results from humans would be very similar to those of Chapter 5 and largely invariant across individuals. In order to more objectively support this claim, I set about demonstrating it by devising a simple psychological experiment that would ask voluntary test subjects to judge what the composing parts of the simple objects used in this thesis were and then trying to draw comparisons between the data collected and the part segmentation produced as in Chapter 5.

Of course, the method presented in this thesis to perform part segmentation from edge images employs a blob or sausage-like model that has been directly inherited from

previous research in the field. On the other hand, in the psychological experiment that is going to be described, we do have objects mostly composed by sausage-like parts but *the test subjects are not explicitly told what kind of model to use*. In this way the scope of the experiment is broader than that of just finding the best combination of part models that fit to the images. Here, people have to first create their mental picture of what a part is *in general* and then apply it to other cases, without the definition being explicitly stated. However, in order for the experiment to be conducted properly, it is necessary that the subjects know what to do with the test images and therefore some initial help must be provided both in terms of linguistic description and pictorial examples.

Hence, differently from the thesis, the present experiment is about *modelling and segmentation fused in a descriptive definition of what the deep nature of an object part might be*.

Before starting the description of the experimental set up and results, I wish to point out that, since I have no formal training in experimental psychology, the spirit of this endeavour is of a certain do-it-yourself nature in that the rules employed for the experimental design and the evaluation of the data collected followed some simple but precise common sense practices.

E.2 The experimental set-up

This section describes the set-up used to carry out the psychological experiment on a small group of voluntary test subjects.

In designing the experiment, two problems had to be faced:

a. How can we give a definition of part that is as much as possible detached from what I – as a person who has already investigated the problem – think of it, which might have lead to biased results. In particular, I explicitly wanted to avoid any kind of technical or mathematical definition (such as Hoffman and Richards' transversality principle [Hoffman & Richards 85]) in favour of a more intuitive, possibly functional definition.

b. How to render the experiment, comprising learning the definition and actual execution, easy and quick enough in order not to put off people who had good intentions of doing it in the first place.

Serious care was taken to the end of satisfying these two requirements, mainly dictated by common sense. During this process, three key design decisions were made, which are discussed below:

The definition of part. The definition provided to the test subjects is given in the Roman-numbered item list of Figure E.2. The first thing to do is to separate objects; this is a fundamental step without which it would have been unclear how to tell people to consider, for instance, the beer bottle body in Image #3

of Figure E.3 should be considered as a single part of several. Then, for each of these distinct objects, the composing parts have to be found following the very qualitative guidelines given in the description: *"Sketch your best guess at the fewest and simplest parts it:* (1) *might be composed of or* (2) *might be made from or* (3) *most easily be broken into"*. As we shall see later, this definition (as would have any other definition), shaped some of the results produced by some test subjects.

The provision of examples. Figures E.4 and E.5 show the examples that has been provided to the test subjects to visually describe what part segmentation is meant to be. The objects were chosen to have an intentionally clear-cut part structure in order to make sure that the test subjects create a mental picture of the problem to go alongside the definition by words given above. This point is arguable because it might be seen as giving too much bias to what the final results should look like in the general case. Despite this, after preliminary trials, it become clear that without these concrete examples the test subjects would have probably not been able to properly execute the experiment. However, the examples shown to the test subjects were simple and with reasonably obvious natural parts (at least of the sort considered in this thesis), whereas the actual examples were somewhat less obvious.

The way results should be produced. Since the domain of parts dealt with in this thesis is a simple one, the most straightforward way of producing experimental data was to sketch object parts in terms of blob-like shapes in their original relative position[1].

After all these key issues were agreed upon the experimental set-up was frozen and an hypertext HTML page was set up providing guidelines for executing the experiment, the test input images and some simple examples of how the final results should look like. All the voluntary test subjects used this page as *sole* input to execute the experiment. Figure E.1 shows an hard copy of this WEB page (viewed through Netscape^{TM}) whereas Figures E.2, E.3, E.4 and E.5 describe the experiment guidelines in a more readable format.

[1] Some people, however, outlined parts on the original edge images themselves either by redrawing or by printing them out first.

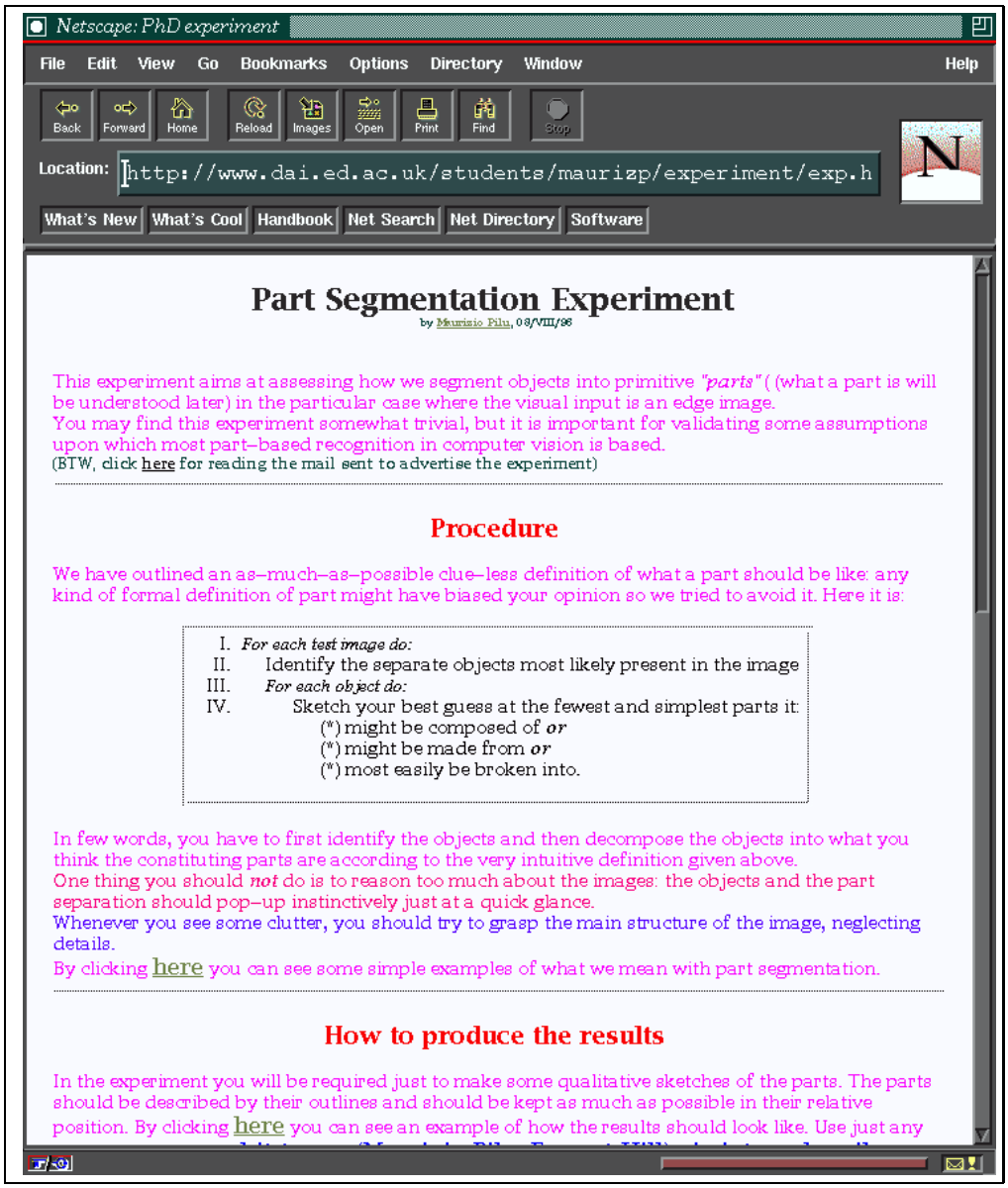


Figure E.1: Screen hard-copy of the HTML page set up to describe the experiments. All the voluntary test subjects used this page as sole input to execute the experiment. The complete description appears in a more readable format next in Figs. E.2, E.3, E.4 and E.5.

Introduction

This experiment aims at assessing how we segment objects into primitive "parts" (what a part is will be understood later) in the particular case where the visual input is an edge image. You may find this experiment somewhat trivial, but it is important for validating some assumptions upon which most part-based recognition in computer vision is based.

Procedure

We have outlined an as-much-as-possible clue-less definition of what a part should be like: any kind of formal definition of part might have biased your opinion so we tried to avoid it. Here it is:

I. For each test image do:
II. Identify the separate objects most likely present in the image
III. For each object do:
IV. Sketch your best guess at the fewest and simplest parts it:
   - might be composed of or
   - might be made from or
   - most easily be broken into.

In few words, you have to first identify the objects and then decompose the objects into what you think the constituting parts are according to the very intuitive definition given above. One thing you should not do is to reason too much about the images: the objects and the part separation should pop-up instinctively just at a quick glance. Whenever you see some clutter, you should try to grasp the main structure of the image, neglecting details. By clicking here (see Figure E.4) you can see some simple examples of what we mean with part segmentation.

How to produce the results

In the experiment you will be required just to make some qualitative sketches of the parts. The parts should be described by their outlines and should be kept as much as possible in their relative position. By clicking here (see Figure E.5) you can see an example of how the results should look like. Use just any paper and then send it to me (Maurizio Pilu, Forrest Hill) via internal mail. The whole experiment should not take no more than a few minutes. Below (see Figure E.3) you can find the six test edge images with which the experiment is to be performed.

Figure E.2: Guidelines for the psychological experiment. It first gives a brief introduction and then describes the procedure for executing the experiment. Finally, some notes are added for helping the test subject to sketch his/her judgements in a coherent and readable way. Note that in the real set-up this was a HTML page and therefore the underlined "here"s were actual hyper-links to the examples.

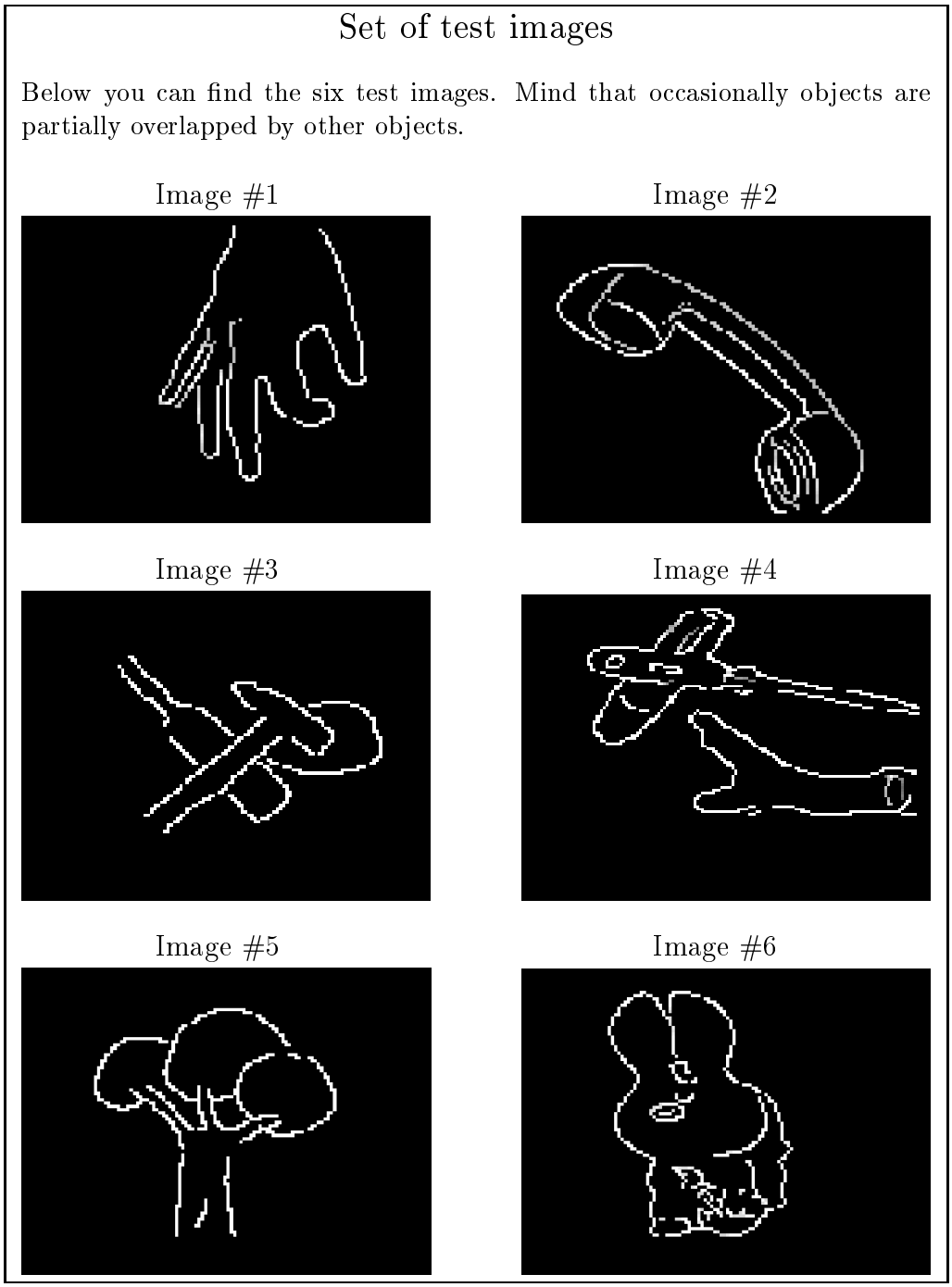


Figure E.3: The six test images on which the experiments was executed. These images were attached to the bottom of the screen corresponding to Figure E.2 in the actual HTML page.

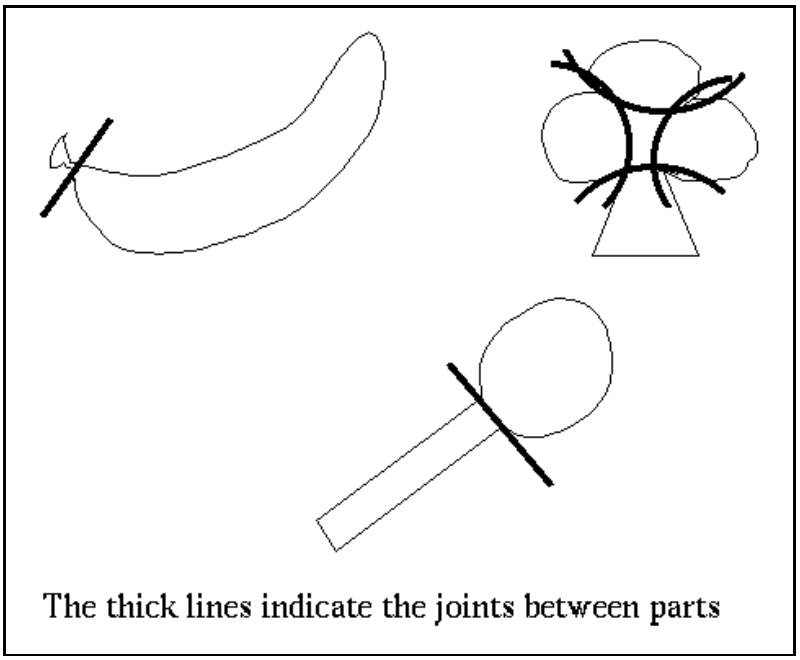


Figure E.4: Example of what part segmentation is meant to be. The objects chosen have an intentionally clear-cut part structure in order to make sure that the test subjects create a mental picture of the problem to go alongside the definition by words given in Figure E.2.

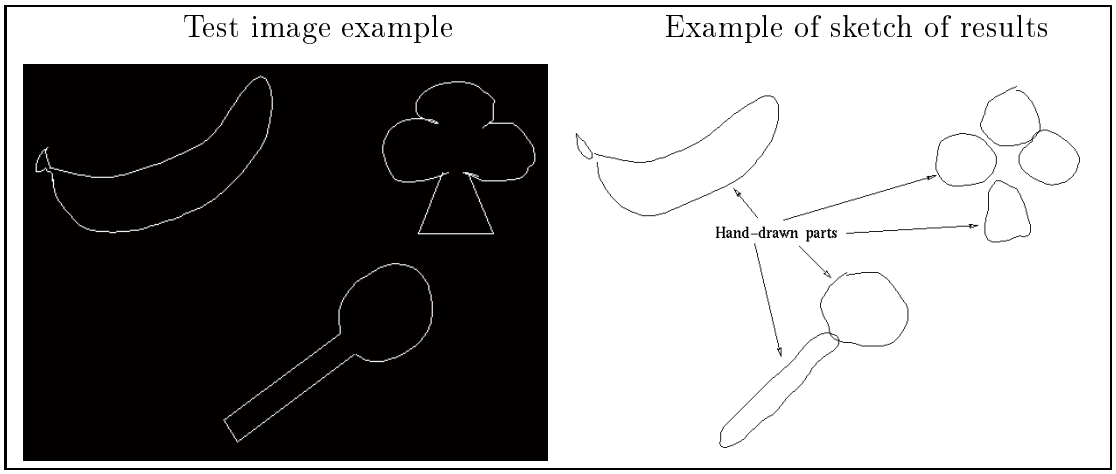


Figure E.5: Left: an example of real input (still example objects); Right: how to produce the results in terms of part blobs. After preliminary trials, it become clear that without these two concrete examples the test subjects would have probably not been able to easily understand the experiment.

E.3 Experimental data collected

This section presents the experimental data collected, the way they have been processed and my interpretation of them.

Eighteen volunteers participated to the experiment. About 15 of them are students or academics working at the Department of Artificial Intelligence.

Following the guidelines, the test subjects sketched the objects and the parts composing them on paper. Some of them took extreme care in reproducing the exact shape of the parts, whereas some others just used blobs (possibly with some sort of deformation). Parts were placed in their original position by most subjects but some sketched the separate objects in different positions.

Once the body of experimental data was gathered, the responses of all subjects for each test image were first divided into classes. New classes were formed whenever a response differed both in terms of number of parts and their relative positioning from others; variations in the parts' dimensions and/or shape were not taken in to consideration because that can be attributed to the different drawing style and/or care with which the experiment was carried out by each test subject.

For convenience, the classification of responses for Image #4 was split for the two sub-images containing the two clearly distinct clusters of edges in the top half (Image #4a) and in the bottom half (Image #4b); having not proceeded in this way would have lead to far too many classes. However the test subjects examined the whole image and, as we shall see in a moment, this has affected both interpretations.

Figure E.6 through Figure E.12 show, image by image, my hand-drawings of the classes of responses that were obtained from all subjects. The classes summarise the individual structures extracted by the test subjects. Each class is tagged by a capital letter and all its parts are numbered. In some cases the subject failed to produce an answer complying to the guidelines and have been declared as "NULL".

Table E.2 gives the overall counts of each class of each image; the individual responses are shown in Table E.1. By looking at the two tables, it appears clear the contrast between some classes that correspond to just one answer by a single subject, and some others that have been particularly popular choices. On the other hand, for some images there have been many different interpretations and for some others the range of choices has been limited. Indeed, the extreme case is Image #5, for which all test subjects agreed on one single rendering. In the following the results are discussed image by image.

In the case of the hand in Image #1 (Figure E.6), class B was the top choice as expected. Note the slight difference between B and D, where the thumb is of different length, and E, where some shading edges near the little finger has made one test subject to draw an additional finger. Perhaps case C should have been considered as NULL but it was classified normally because interestingly a test subject used his high level knowledge of the fingers' bone structure.

Expectedly, for the handset in Image #2 (Figure E.7), class B was the most popular response. The classes A, D and to some extent E have got the small parts A/3, D/3

INDIVIDUAL RESPONSES

	Im #1	Im #2	Im #3	Im #4/a	Im #4/b	Im #5	Im #6
Subject #1	E	C	A	B	B	A	C
Subject #2	A	B	A	C	C	A	I
Subject #3	B	B	A	B	C	A	H
Subject #4	B	B	B	B	F	A	I
Subject #5	B	B	A	B	A	A	D
Subject #6	B	D	A	D	A	A	NULL
Subject #7	B	B	A	B	A	A	G
Subject #8	C	C	C	F	G	A	F
Subject #9	D	A	A	E	E	A	B
Subject #10	B	D	A	B	C	A	D
Subject #11	A	B	A	C	D	A	E
Subject #12	B	B	A	B	C	A	J
Subject #13	B	NULL	A	B	B	A	C
Subject #14	B	B	A	B	A	A	B
Subject #15	B	E	A	A	C	A	A
Subject #16	B	B	A	C	C	A	I
Subject #17	B	D	A	B	G	A	J
Subject #18	A	C	A	B	E	A	C

Table E.1: Experimental data collected. It can be noticed the contrast between some classes that correspond to just one answer by a single subject, and some others that have been particularly popular choices. Refer to Figure E.6 through Figure E.12 for looking up tag letters to the classified responses.

and E/3 which are meant to be the microphone and speaker covers and the subjects probably used their high-level knowledge of an old-style handset mechanical structure. Perhaps, class C and should have also been considered NULL, if it did not come from two different subjects (one of whom, Subject #8, also produced Image #1/C). Curiously, in class E the absence of a clear edge in Image #2 between the handle and the top piece made a subject perceive a squash-like shape.

In the relatively easy case of Image #3 (Figure E.8), class A was overwhelmingly the most frequently chosen answer. Despite that, a subject saw B/2 and B/3 as disjoint; another one saw the hammer head as composed of two parts (C/5 and C/6) and also the bottle neck and bottle body separated by a sort of "shoulder" (C/2).

Image #4a (Figure E.9) received a funny interpretation by most subjects. Rather that seeing a screw driver and another less clear object beneath (actually a marker), they saw a small airplane and the screw-driver shaft was interpreted as its smoke trail or a banner. Because of this, the most popular response was B, in which the two alleged wings are considered two separate objects. In A and D the shaft was not reported, perhaps because of its thinness. Case C, which detected a single object (which was actually the case in the original scene) beneath the screw-driver handle, was selected by just a handful of subjects. Worth noticing is also case F, where the little part F/5 was included by the same Subject #8 that reported part C/2 in Image #3.

CLASSES COUNT

	A	B	C	D	E	F	G	H	I	J	NULL
Im #1	3	12	1	1	1						
Im #2	1	9	3	3	1						1
Im #3	16	1	1								
Im #4/a	1	11	3	1	1	1					
Im #4/b	4	2	6	1	2	1	2				
Im #5	18										
Im #6	1	2	3	2	1	1	1	1	3	2	1

Table E.2: Classes count: the different classes of interpretation are counted for each image. For some images there have been many different interpretations and for some others the range has been limited indeed, the extreme case being Image #5, for which all test subjects agreed on one single rendering. Refer to Figure E.6 through Figures E.12 for looking up tag letters to the classified responses.

Results are a quite interesting for Image #4b (Figure E.10) too. The four test subjects that decided for the single-part class A also had the "airplane" interpretation of Image #4a because (they were asked about their choice) they thought it was a cloud; this was quite surprising because although the image hardly resembles any kind of cloud-like shape I have ever observed, the sky scenario that some people imagined took over a more rational interpretation. Even, a subject saw the tail of a fighter plane in it and another one saw a hand throwing a toy plane! Beside that, apart from class G which came again by the over-detailing Subject #8 and the little details D/2 and E/4, the three other meaningful classes of responses are B, C and F, which are, in my opinion, equally good interpretations differing only in the almost arbitrary choice on where and how big the main body is.

In the case of the tree in Image #5 (Figure E.11), the response was unique. People interpreted it unambiguously and probably the distinct part structure of it popped up at a first glance. We are all familiar with tearing off small branches or shearing bushes and probably this strong imagery of *what you can do* with a tree determined this clear kind of common response.

Finally, the case of Image #6 (Figure E.12) turned out to be quite a hard one to classify. Most people correctly saw a toy rabbit in the image, most probably helped by the two characteristic big ears. In all responses (except E) head and ears were clearly identified. Curiously enough, the nose was always reported. Regarding the lower part of the body, the cluttering and the side shadow edge caused a multiplicity of interpretations, ranging from two legs without a body (B,C,H), a big body and legs (A), body and paws (I,F), and so forth. Probably, there is no point in trying to speculate upon the reasons why people gave some many disparate answers for the lower body, because the quality of the edge image there was really appalling. The most amazing interpretation (which I considered as "NULL" for its weirdness) is that Image #6 corresponded to two gnomes cuddling each other, a picture that pops immediately up after one is told, as in any good optical illusion.

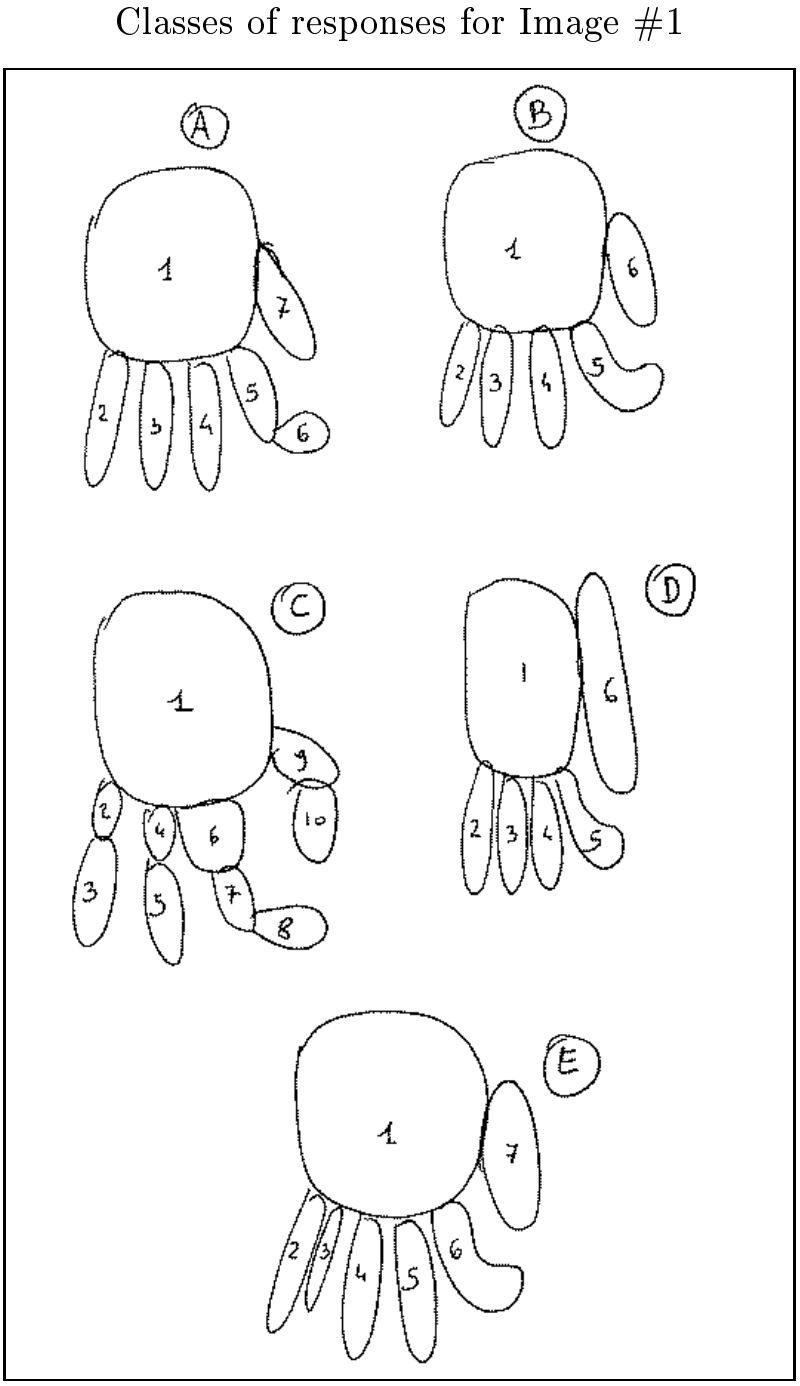


Figure E.6: Classes of responses for Image #1. Class B was the top choice as expected. Note the slight difference between B and D, where the thumb is of different length, and E, where some shading edges near the little finger has made one test subject to draw an additional finger. Perhaps case C should have been considered as NULL but it was classified normally because interestingly a test subject used his high level knowledge of the fingers' bone structure.

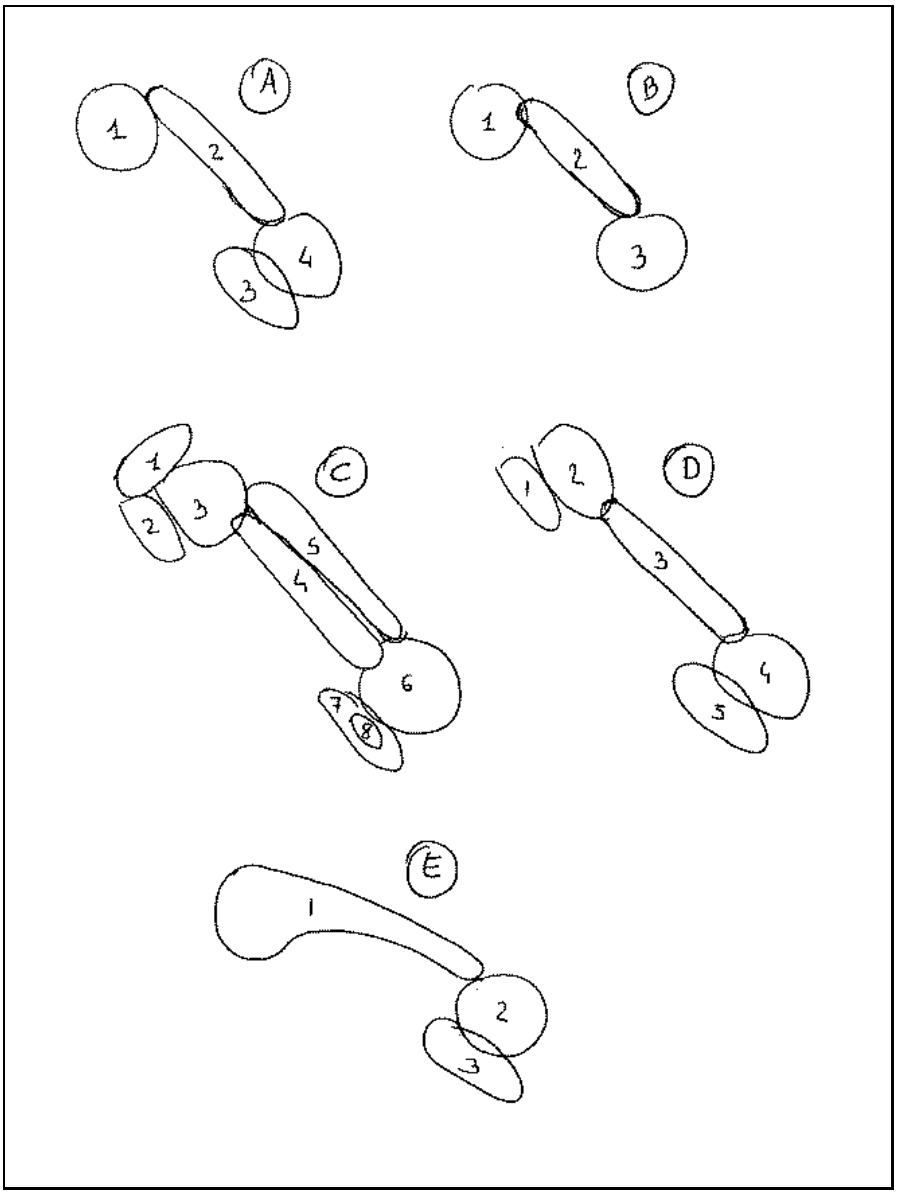


Figure E.7: Classes of responses for Image #2. Expectedly class B was the most popular response. The classes A, D and to some extent E have got the small parts A/3, D/3 and E/3 which are meant to be the microphone and speaker covers and the subjects probably used their high-level knowledge of an old-style handset mechanical structure. Perhaps, class C and should have also been considered NULL, if it did not come from two different subjects (one of whom, Subject #8, also produced Image #1/C). Curiously, in class E the absence of a clear edge in Image #2 between the handle and the top piece made a subject perceive a squash-like shape.

Classes of responses for Image #3

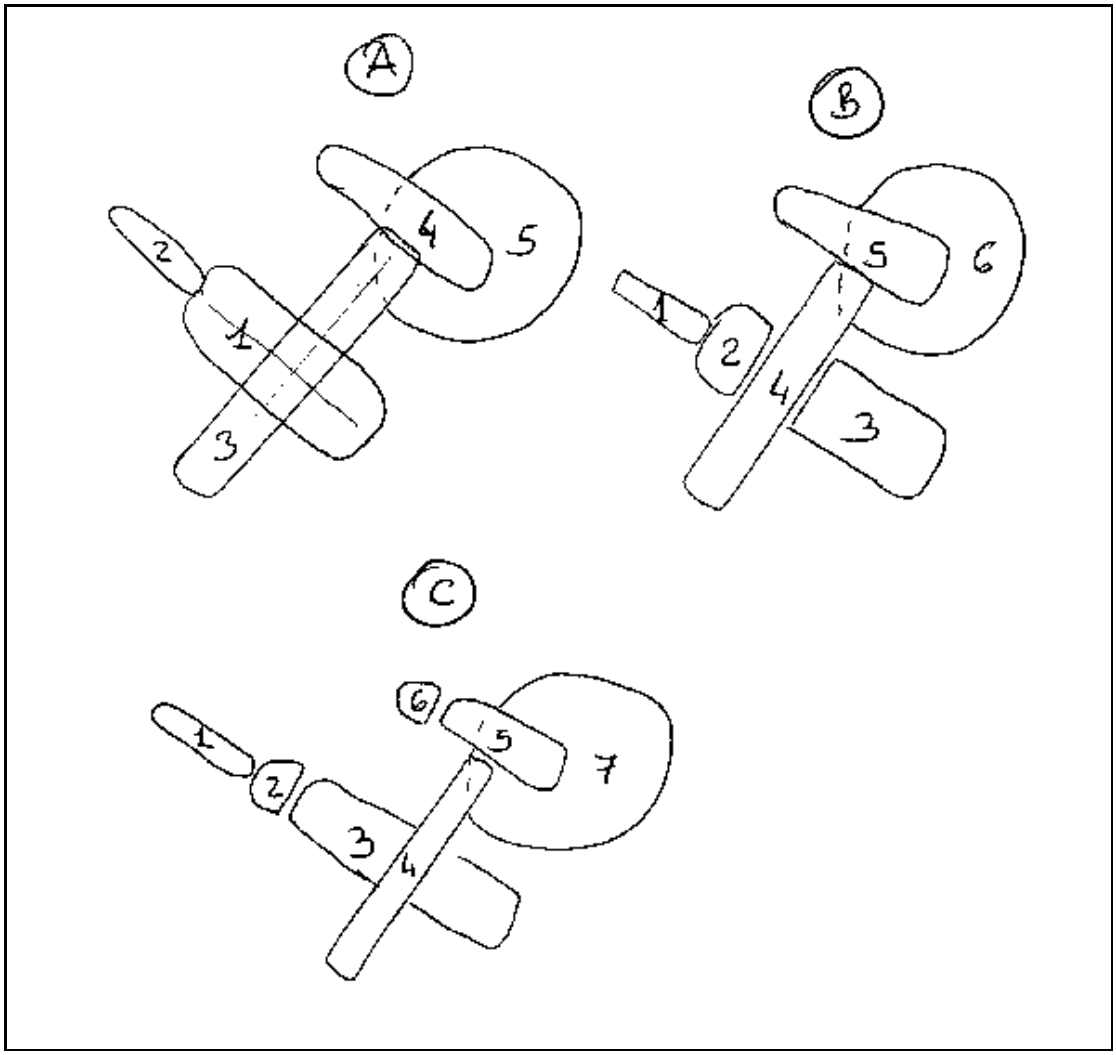


Figure E.8: Classes of responses for Image #3. class A was overwhelmingly the most frequently chosen answer. Despite that, one subject saw B/2 and B/3 as disjoint; another one saw the hammer head as composed of two parts (C/5 and C/6) and also the bottle neck and bottle body separated by a sort of "shoulder" (C/2).

Classes of responses for Image #4a

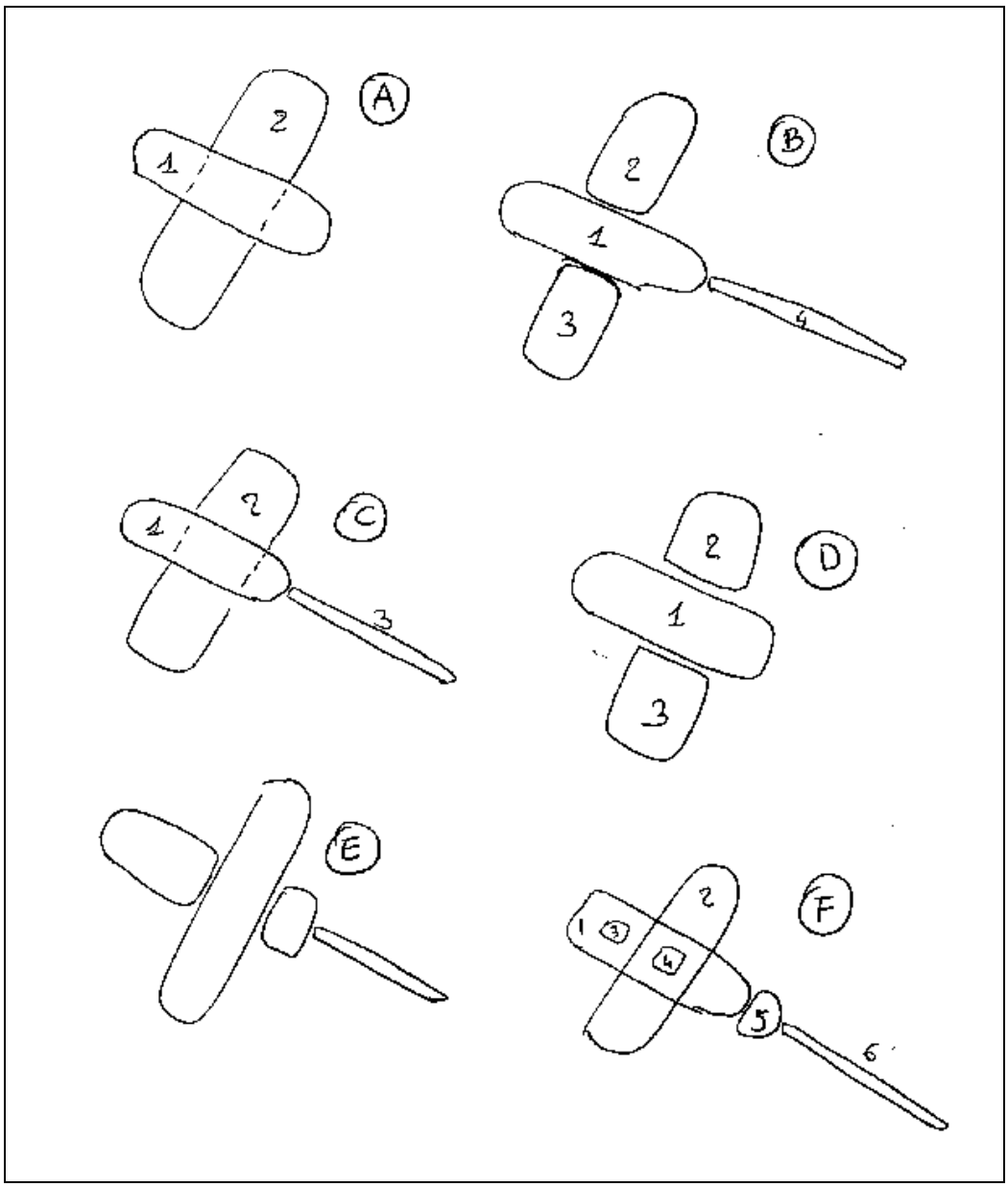


Figure E.9: Classes of responses for Image #4a. The cluster of edges in the top half of image #4 received a funny interpretation by most subjects. Rather that seeing a screw driver and another less clear object beneath (actually a marker pen), they saw a small airplane and the screw-driver shaft was interpreted as its smoke trail or a banner. Because of this, the most popular response was B, in which the two alleged wings are considered two separate objects. In A and D the shaft was not reported, perhaps because of its thinness. Case C, which detected a single object (which was actually the case in the original scene) beneath the screw-driver handle, was selected by just a handful of subjects. Worth noticing is also case F, where the little part F/5 was included by the same Subject #8 that reported part C/2 in Image #3.

Classes of responses for Image #4b

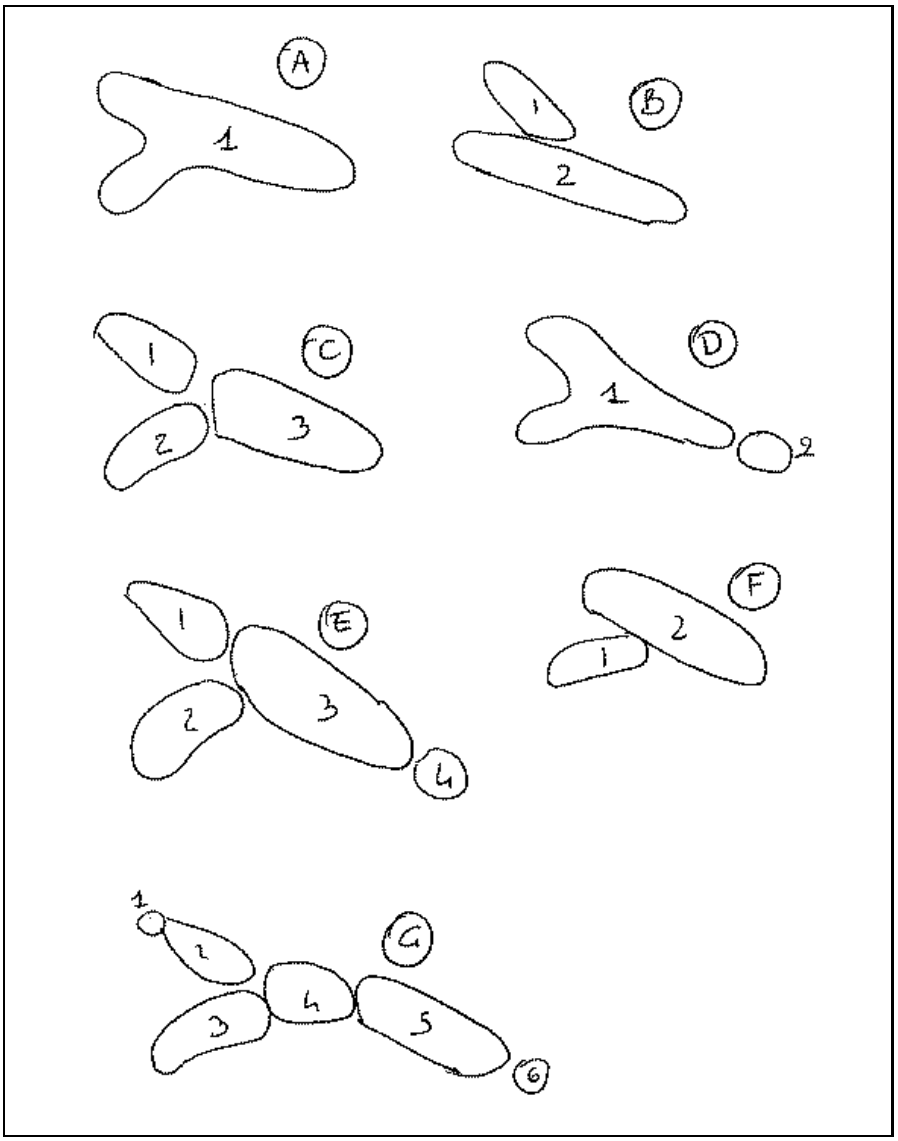


Figure E.10: Classes of responses for Image #4b. Results here are quite interesting too. The four test subjects that decided for the single-part class A also had the "airplane" interpretation of Image #4a because (they were asked about their choice) they thought it was a cloud; this was quite surprising because although the image hardly resembles any kind of cloud-like shape I have ever observed, the sky scenario people imagined took over a more rational interpretation. Even, a subject saw the tail of a fighter plane in it and another one saw a hand throwing a toy plane! Beside that, apart from class G which came again by the over-detailing Subject #8 and the little details D/2 and E/4, the three other meaningful classes of responses are B, C and F, which are in my opinion equally good interpretations differing only in the almost arbitrary choice on where and how big the main body is.

Classes of responses for Image #5

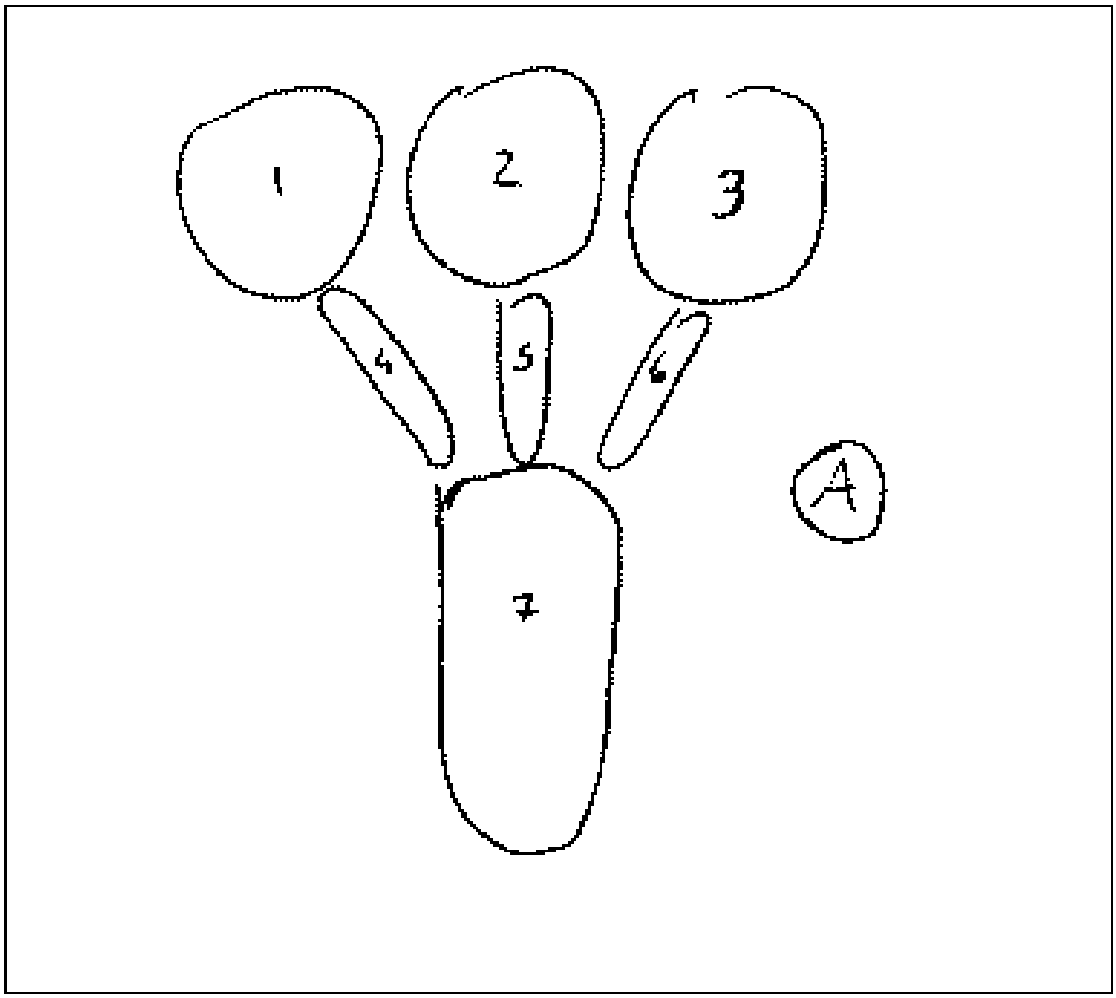


Figure E.11: Classes of responses for Image #5. In this case the answer was unanimous. People interpreted the tree image unambiguously and probably the distinct part structure of it popped up at a first glance. We are all familiar with tearing off small branches or shearing bushes and probably this strong imagery of *what you can do* with a tree determined this clear kind of common response.

Classes of responses for Image #6

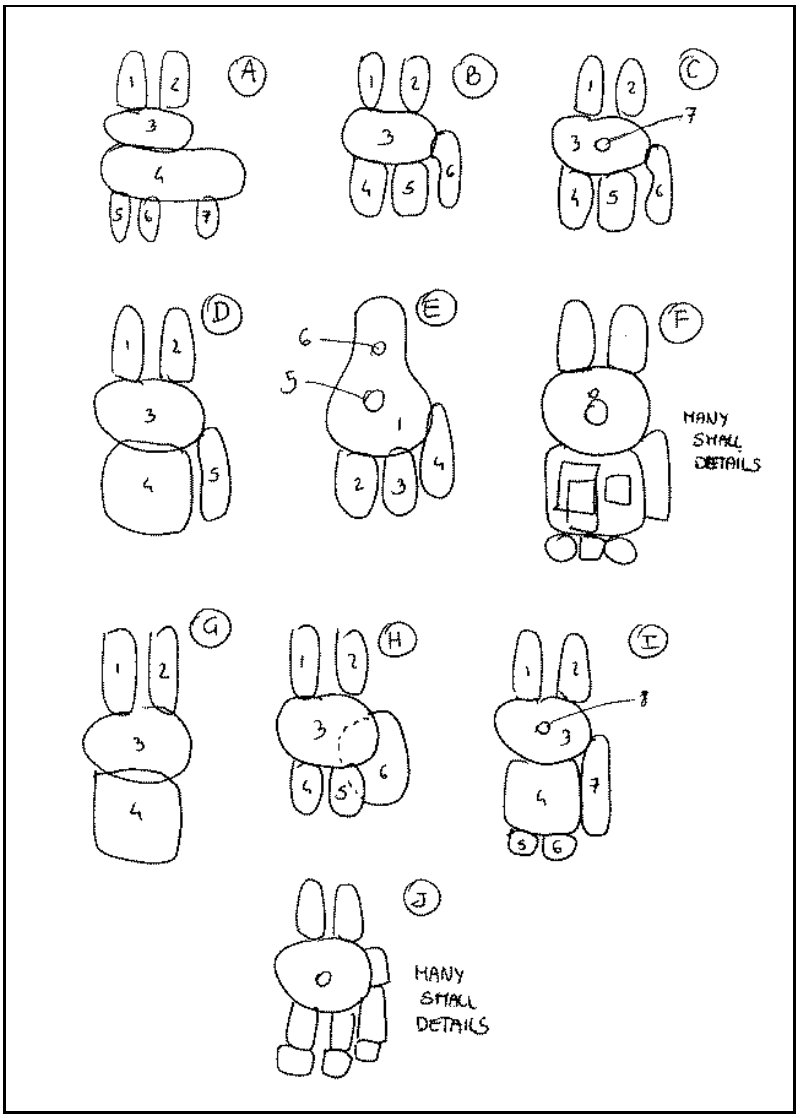


Figure E.12: Classes of responses for Image #6. This case turned out to be quite a hard one to classify. Most people correctly saw a toy rabbit in the image, most probably helped by the two characteristic big ears. In all responses (except E) head and ears were clearly identified. Curiously enough, the nose was not always reported. Regarding the lower part of the body, the cluttering and the side shadow edge caused a multiplicity of interpretations, ranging from two legs without a body (B,C,H), a big body and legs (A), body and paws (I,F), and so forth. Probably, there is no point in trying to speculate upon the reasons why people gave so many disparate answers for the lower body, because the quality of the edge image there was really appalling. The most amazing interpretation (which I considered as "NULL" for its weirdness) is that Image #6 corresponded to two gnomes cuddling each other, a picture that pops immediately up after one is told, as in any good optical illusion.

E.4 Discussion

In this section I will endeavour to make a connection between the data collected by the psychological experiment just described with the part segmentation results obtained throughout this thesis. In particular, because of the nature of the experiment, comparisons will be drawn between the classes of responses and the output of the Minimum Description Length part filtering method presented in Chapter 5.

Figure E.13 reproduces here for convenience the part segmentation results that will be used for comparisons and will be referred to in the rest of the section. These images have been placed following the order of Figure E.3 and not that used in Chap. 5. Note that this representative outputs have been chosen using a single parameter configuration (see Sec. 5.3.6 and Sec. 5.3.7).

For Image #1 the MDL results are comparable to classes A or B in Figure E.6. However, for reasons well explained in Sections 5.3.6/ 5.3.7 the last segment of the index does not appear in the MDL output and therefore it is somehow in between classes A and B; however, considering that classes A and B account for 15 out of 18 responses, this can be seen as a good result. Notice that the back of the hand, as said in 5.3.6 could not be recovered by the MDL filtering because the correct hypothesis was not generated. Finally, the second "longer" interpretation for the thumb in the MDL output has also a timid correspondence in the single answer of class D in Figure E.6.

In the case of the handset in Image #2, the MDL results are in harmony with those represented by classes A, B and D of Figure E.7 but as in the previous case, they do not match any of them precisely. Rather, the MDL output has three parts of class B (B/1, B/2 and B/3) but also the "covers" as in A/3, D/1 and D/5. The spurious interpretation of the handle appears also in class C, albeit differently. Classes A, B and D account for 13 of the 17 valid responses and therefore this MDL result can be considered positive too.

For the composite Image #3 where actual objects were overlapping, the (very stable) MDL result is precisely that of class A in Figure E.8, which account for 16 responses out of 18. By any standards, therefore, the MDL has performed excellently.

For the top half of Image #4, named Image #4a in the previous section, the results are not so exciting ... *for the test subjects*. Because of the high level interpretation given to the image, the test subjects separated the occluded object slanted by 45 degree (actually a marker) in two halves corresponding to the two wings of an airplane (see Figure E.9 and comments therein). The "dull" MDL filtering, on the other hand, saw that the one-model interpretation was cheaper than the two-model one and selected that one. The one-object interpretation is the correct one and had not the test subjects reasoned too much about the meaning of that poor-quality image (or should have the experiment been formulated differently?), perhaps they would have given the correct, more low-level interpretation.

As said in the previous section, for Image #4b (bottom half of Image #4), the results of the psychological experiment shown in Figure E.10 have been affected by (see caption of Figure E.10) two factors: *a)* scenario imagined and *b)* the *inherently arbitrary decomposition*. In the case of the MDL output, the result has been conditioned by

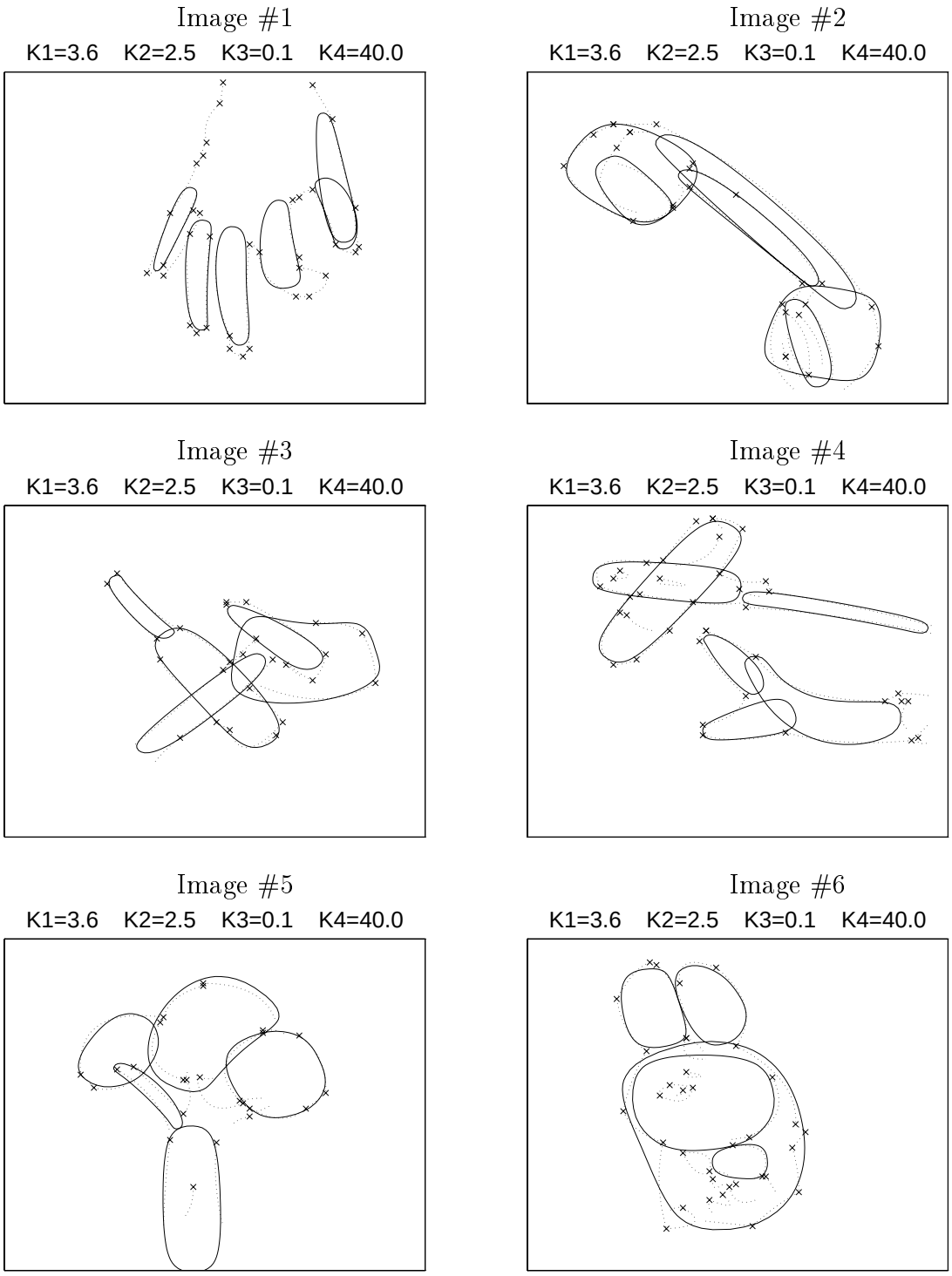


Figure E.13: Results produced by the MDL filtering method of Chapter 5. Note that this representative outputs have been chosen using a single parameter configuration (see Sec. 5.3.6 and Sec. 5.3.7). These images have been placed following the order of Figures E.3 and not that used in Chap. 5.

the non-creation of the two elongated hypotheses encompassing the whole length of the object prolonging in the top and bottom branches, respectively; because of this, in fact, hypotheses like B and F in Figure E.10 could not possibly happen. Obviously, because of the nature of the models employed, also classes A and D could not occur but that is not a big problem because they can be considered as failed decomposition by "too holistic" subjects. To come down to numbers, by considering classes C and E together (they just differ for the little tail part), the MDL output is in accordance to 8 out of 18 answers.

The tree in Image #5 has received a single interpretation by all test subjects in the psychological experiment, which is shown in Figure E.11. As pointed out in Section 5.3.6, in the MDL output the middle and right little branches have not been selected purely for scale reasons; if this detail is not taken into account, the results are in perfect correspondence.

Finally, there is the messy case of the toy rabbit of Image #6. The interpretation produced by the MDL filtering method is quite simple, as it can be seen in Figure E.13. If we neglect the illegible lower body of the toy rabbit, it turns out that ears and head, the most evident entities, have been perceived by 16 out of 17 test subjects in the psychological experiment, as seen in Figure E.12; happily, these three parts appear stably also in the MDL output. For the lower part of the body the test subjects have given a plethora different interpretations (see caption in Figure E.12) but the MDL did not perform better either, as described in the captions of Figures 5.31 and 5.31.

Summing up, overall (and whenever possible) the part segmentation produced by the method described in this thesis does produce results in accordance to those given by many of the test subjects. The test objects used in the test scenes were of limited complexity because of the inherent limitation of part-based representation schemes and for some other limits and pitfalls in the techniques used, but the test subjects were not told what kind of model to use in their decomposition so the experiment can be considered fair from this point of view.

It was a rather weird experience to compare a computer program and humans for such an *under-specified task* as that in question, since everybody knows that it is pointless at this technological stage: on one side a still embryonic computer vision technique and on the other a magnificent thinking machine as a person. And yet these results, in their surreal, absurd simplicity are comparable. Perhaps the main conclusion (and original aim) of the experiment is the confirmation that part decomposition is not so much an under-specified task after all.

Bibliography

[Bajcsy & Solina 87] R. Bajcsy and F. Solina. Three dimensional object representation revisited. In *Proceedings of the International Conference of Computer Vision*, pages 231–240, 1987.

[Barr 81] A.H. Barr. Superquadrics and angle-preserving transformations. *IEEE Computer Graphics and Applications*, 1(1):11–23, January 1981.

[Barrow & Tenenbaum 81] H.G. Barrow and J.M. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17:75–116, 1981.

[Bennamoun & Boashash 94] M. Bennamoun and B. Boashash. Integration of a part segmentation based vision system. In *Proceedings of the IEEE International Conference on Image Processing*, volume 3, pages 513–517, Austin, TX, November 1994.

[Bergevin & Levine 93] R. Bergevin and M. Levine. Generic object recognition: Building and matching coarse description from line drawings. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(1):19–36, January 1993.

[Bergevin 90] R. Bergevin. Primal access recognition of visual objects. Technical Report TR-CIM-90-5, Mc Gill University, Canada, February 1990.

[Biederman & Gerhardstein 95] I. Biederman and P.C. Gerhardstein. Viewpoint-dependent mechanisms in visual object recognition - Reply to Tarr and Bulthoff (1995). *Journal of Experimental Psychology - Human Perception and Performance*, 21(6):1506–1514, 1995.

[Biederman 87] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[Binford 71] T.O. Binford. Visual perception by computer. In *Proceedings of the IEEE System Science and Cybernetic Conference*, Miami, December 1971.

[Blake & Zisserman 87] A. Blake and A. Zisserman. *Visual Reconstruction.* MIT Press, Cambridge, MA, 1987.

[Blum & Nagel 78] H Blum and R.N. Nagel. Shape description using weighted symmetry axis features. *Pattern Recognition*, 10:167–180, 1978.

[Bookstein 79] F.L. Bookstein. Fitting conic sections to scattered data. *Computer Graphics and Image Processing*, (9):56–71, 1979.

[Borges 96] D.L Borges. *Recognizing Three-Dimensional Objects using Parametrized Volumentric Models.* Unpublished PhD Thesis, Department of Artificial Intelligence, University of Edinburgh, May 1996.

[Brady 87] Michael Brady. Seeds of perception. In *Proceeding of the ALVEY Conference*, pages 259–265, 1987.

[Brooks 84] R. Brooks. *Model Based Computer Vision.* UMI Research Press, MI, USA, 1984.

[Burns *et al.* 94] J.B. Burns, H.K. Nishihara, and S.J. Rosenschein. Appropriate-scale local centers: A foundation to part based recognition. In *Proceedings of the ARPA Image Understanding Workshop*, pages 1281–1286, Monterey. CA, November 1994.

[Checkosky & Whitlock 73] S.F. Checkosky and D. Whitlock. Effects of pattern goodness on recognition time in a memory search task. *Journal of Experimental Psychology*, 100:341–348, 1973.

[Chen & Kak 89] C. Chen and A. Kak. A robot vision system for recognizing 3D objects in low-order polynomial time. *IEEE Transaction Syst. Man and Cybern.*, 19:1535–1563, 1989.

[Clowes 65] M. Clowes. On seeing things. *Artificial Intelligence*, 2:79–116, 1965.

[Cootes & Taylor 92] T.F. Cootes and C.J. Taylor. Active shape models - 'smart snakes'. In *Proceedings of the British Machine Vision Conference*, pages 266–275, 1992.

[Cootes & Taylor 94] T.F. Cootes and C.J. Taylor. Combining point distribution models with shape models based on finite element analysis. In *Proceedings of the British Machine Vision Conference*, pages 419–428, 1994.

[Cootes & Taylor 95] T.F. Cootes and C.J. Taylor. Active shape models: A review of recent research. In K.V. Mardia and C.A. Gill, editors, *Proceedings in Current Issues in Statisctical Shape Analysis*, pages 108–114, Leeds, April 1995.

[Cootes & Taylor 96] T.F. Cootes and C.J. Taylor. Locating objects of varying shape using statistical feature detectors. In *Proceedings of the European Conference on Computer Vision*, pages 465–474, 1996.

[Cootes *et al.* 91] T.F. Cootes, D. Cooper, C.J. Taylor, and J. Graham. A trainable method of parametric shape description. In *Proceedings of the British Machine Vision Conference*, pages 54–61. Springer-Verlag, 1991.

[Cootes *et al.* 94] T.F. Cootes, A. Hill, C.J. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–365, 1994.

[Cox *et al.* 92] I. Cox, J Rehg, and S. Hingorani. A Bayesian multiple hypotheses approach to contour grouping. In *Proceedings of the European Conference on Computer Vision*, pages 72–77, 1992.

[Critton & Parish 83] C.W.K. Critton and E.A. Parish. Boundary location from an initial plan: The bead chain alogrithm. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 5(1):8–14, January 1983.

[Darrell & Pentland 95] T. Darrell and A.P. Pentland. Cooperative robust estimation using layers of support. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(5):474–487, May 1995.

[Dickinson *et al.* 92a] S. J. Dickinson, A.P. Pentland, and A. Rosenfeld. From volumes to views: An approach to 3D object recognition. *Computer Vision, Graphics and Image Processing*, 55(2):130–154, 1992.

[Dickinson *et al.* 92b] S.J. Dickinson, A.P. Pentland, and A. Rosenfeld. 3D Shape Recovery Using Distributed Aspect Matching. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 14(2):130–154, 1992.

[Dickinson *et al.* 93] S.J. Dickinson, R. Bergevin, I. Biederman, J.O. Eklundh, R. Munck-Fairwood, and A. Pentland. The use of geons for generic 3D object recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1693–1699, Chambery, France, 1993.

[Eggert & Bowyer 93] D. Eggert and K. Bowyer. Computing the perspective projection aspect graph of solids of revolution. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(2):109–128, 1993.

[Eggert *et al.* 95] D. Eggert, L. Stark, and K. Bowyer. Aspect graphs and their use in object recognition. *Annals of Mathematics and Artificial Intelligence*, 13:347–375, 1995.

[Ellis *et al.* 92] T. Ellis, A. Abbood, and B. Brillault. Ellipse detection and matching with uncertainty. *Image and Vision Computing*, 10(2):271–276, 1992.

[Ferryman *et al.* 95] J.M. Ferryman, A.D. Worrall, G.D. Sullivan, and K.D. Backer. A generic deformable model for vehicle recognition. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 127–136, Birmingham, September 1995.

[Fischler & R.C.Bolles 86] M.A. Fischler and R.C.Bolles. Perceptual organization and curve partitioning. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 8(1):100–105, 1986.

[Fitzgibbon & Fisher 92] A. W. Fitzgibbon and R. B. Fisher. Practical aspect graph derivation incorporating feature segmentation performan ce. In *Proceedings of the British Machine Vision Conference*, pages 580–589, September 1992.

[Fitzgibbon & Fisher 95] A.W. Fitzgibbon and R.B. Fisher. A buyer's guide to conic fitting. In *Proceedings of British Machine Vision Conference*, Birmingham, 1995.

[Fitzgibbon 96] A.W. Fitzgibbon. *The representation and extraction of curves.* PhD Thesis, Department of Artificial Intelligence, University of Edinburgh, May 1996. Forthcoming.

[Fitzgibbon *et al.* 96] A. W. Fitzgibbon, M. Pilu, and R.B. Fisher. Direct least squares fitting of ellipses. In *International Conference on Pattern Recognition*, pages xxx–xxx, Vienna, August 1996.

[Franklin & Barr 81] W.R. Franklin and A. Barr. Faster calculation of superquadric shapes. *IEEE Computer Graphics and Applications*, July 1981.

[Freeman 78] H Freeman. Shape description via the use of critical points. *Pattern Recognition*, 10:159–166, 1978.

[Friedland & Rosenfeld 92] N.H. Friedland and A. Rosenfeld. Compact object recognition using energy-function-based optimization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 14(7):770–777, 1992.

[Fua & Hanson 89] P. Fua and A.J. Hanson. Objective functions for feature discrimination: Applications to semiautomated and automated feature extraction. In

DARPA Image Understanding Workshop, pages 676–694, 1989.

[Fua 89] P. Fua. Objective functions for feature discrimination: Theory. In *DARPA Image Undertanding Workshop*, pages 443–460, 1889.

[Fuger *et al.* 94] H. Fuger, G. Stein, and U. Stilla. Multi-population evolution strategies for structural image analysis. In *IEEE Conference on Evolutionary Computation*, volume 1, pages 229–234, Orlando, FL, 1994.

[Gander *et al.* 94] W. Gander, G.H. Golub, and R. Strebel. Least-square fitting of circles and ellipses. *BIT*, (43):558–578, 1994.

[Gardner 65] M. Gardner. The superellipse: a curve that lies between the ellipse and the rectangle. *Scientific American*, September 1965.

[Garner 74] W.R. Garner. *The processing of information and structure*. Wiley, New York, 1974.

[Geman & Geman 84] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[Gibson 79] J.J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

[Gnanadesikan 77] R. Gnanadesikan. *Methods for statistical Data analysis of multivariate observations*. Wiley, New York, 1977.

[Grefensette 96] J. J. Grefensette. Optimization of control parameters for genetic algorithms. *IEEE Transaction on System, Man and Cybernetics*, 16:122–128, 1896.

[Hara *et al.* 92] J. Hara, H. Kato, and S. Inokuchi. Description and recognition of animal shilouette image using ellipsoid-expansion. *Systems and Computers in Japan*, 23(7):66–77, 1992.

[Haralick & Shapiro 92] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley, 1992.

[Harnad 87] S. Harnad. Category induction and categorization. In S. Harnad, editor, *Categorical Perception*. Cambridge University Press, 1987.

[Haslam *et al.* 95] J. Haslam, C.J. Taylor, and T.F. Cootes. A probabilistic fitness measure for deformable template models. In *Proceedings of the British Machine Vision Conference*, pages 33–42, Birmingham, September 1995.

[Hill & Taylor 92] A. Hill and C.J. Taylor. Model-based image interpretation using genetic algorithms. *Image and Vision Computing*, 10(5):295–300, June 1992.

[Hoffman & Richards 85] D. Hoffman and W. Richards. Parts of recognition. In A. Pentland, editor, *From Pixels to Predicates*. Ablex, Norwood, NJ, 1985.

[Horn 89] B.K.P. Horn. Shape from shading: a method for obtaining the shape of a smooth opaque object from one view. In M.J. Brooks and B.P.K. Horn, editors, *Shape from Shading*. MIT Press, 1989.

[Huber 81] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[Huffman 71] D. Huffman. Impossible objects as nonsense sentences. *Machine Intelligence*, 6, 1971.

[Hummel & Biederman 92] J.E. Hummel and I. Biederman. Dynamic binding in a neural net model for shape recognition. *Psychological Review*, 99:480–517, 1992.

[Huttenlocher & Wayner 92] D. Huttenlocher and P. Wayner. Finding convex edge groupings in an image. *International Journal of Computer Vision*, 8(1):7:29, 1992.

[Ikeuchi 87] K. Ikeuchi. Generating an interpretation tree from a CAD model for 3D object recognition. *Proceedings of the International Journal of Computer Vision*, pages 145–165, 1987.

[Ittleson 52] W.H. Ittleson. *The Ames Demonstrations in Perceptions*. Hafner, New York, 1952.

[Jacobs 96] D. W. Jacobs. Robust and efficient detection of convex groups. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(1):23–37, January 1996.

[Joliffe 86] I. T. Joliffe. *Principal Components Analysis*. Springer-Verlag, 1986.

[Kanatani 94] K. Kanatani. Statistical bias of conic fitting and renormalization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 16(3):320–326, 1994.

[Kass *et al.* 88] M. Kass, A. Witkin, and D. Terzopulos. Snakes: active contour models. *International Journal of Computer Vision*, 1:321–331, 1988.

[Kendall 84] D.G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bull. Lond. Math. Society*, 16:81–121, 1984.

[Kender & Freudenstain 87] J.R. Kender and D.G. Freudenstain. What is a "degenerate'' view? In *DARPA Image Understanding Workshops*, pages 589–598, 1987.

[Kimia *et al.* 95] B.B. Kimia, A.R. Tannenbaum, and S.W. Zucker. Shapes, shocks and deformations, I: The components of 2-dimensional shape and the reaction-diffusion space. *International Journal of Computer Vision*, 15(3):189–224, 1995.

[King *et al.* 76] M. King, G. E. Meyer, J. Tangey, and I. Biederman. Shape constancy and a perceptual bias towards symmetry. *Perception & Psychopysics*, 19:129–136, 1976.

[Kirkpatrick *et al.* 83] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[Koenderink & vanDoorn 79] J.J. Koenderink and A.J. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetic*, 32:211–216, 1979.

[Koenderink & vanDoorn 82] J.J. Koenderink and A.J. van Doorn. The shape of smooth objects and the way contours end. *Perception*, 11:129–137, 1982.

[Kohler 59] W. Kohler. *Gestalt Psychology.* Mentor Books, New York, 1959.

[Lakoff 87] G. Lakoff. *Women, Fire and Dangerous Things: what categories reveal about the mind.* The University of Chicago Press, 1987. 2nd edition.

[Leavers 92] V.F. Leavers. *Shape Detection in Computer Vision Using the Hough Transform.* Springer-Verlag, 1992.

[Leclerc 89] Y.G. Leclerc. Constructing simple stable description for image partitioning. *International Journal of Computer Vision*, 3:73–102, 1989.

[Leonardis 93] Ales Leonardis. *Image Analysis using parametric models.* PhD Thesis, University of Ljubljana, 1993. Supervisors: Ruzena Bajcsy and Franc Solina.

[Leonardis *et al.* 90] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation as the best description of the image in term of primitives. In *International Conference on Computer Vision*, pages 121–125, 1990.

[Leonardis *et al.* 94] A. Leonardis, F. Solina, and A. Macerl. A direct recovery of superquadric models in range images using recover-and-select paradigm. In *Proceedings*

of the European Conference on Computer Vision, pages 309–318. Springer-Verlag, 1994.

[Leonardis *et al.* 95] A. Leonardis, A. Gupta, and R. Bajcsy. Segmentation of range images as the search for geometric paramatric models. *International Journal of Computer Vision*, 14:253–277, 1995.

[Leou & Tsai 87] J. J. Leou and W.H. Tsai. Automatic rotational symmetry determination for shape analysis. *Pattern Recognition*, 20:571–582, 1987.

[Les93] Lester Ingber Research, Mc Lean, VA. *Adaptive Simulated Annealing*, 1993. [ftp.alumni.caltech.edu./pub/ingber/ASA.tar.zip].

[Leyton 92] M. Leyton. *Simmetry, Casuality, Mind.* MIT Press, 1992.

[Lim 94] K.G. Lim. *Area Based Stereo.* PhD Thesis, University of Cambridge, April 1994.

[Lindeberg 94] T. Lindeberg. *Scale-Space Theory in Computer Vision.* Kluver, Dordrecht, The Netherlands, 1994.

[Link & Zucker 88] N.K Link and S.W. Zucker. Corner detection in curvilinear dot grouping. *Biological Cybernetics*, 59:247–256, 1988.

[Lowe 85] D. Lowe. *Perceptual Organization and Visual Recognition.* Kluwer Academic Publishers, Boston, MA, 1985.

[Lowe 88] D. Lowe. Organization of smooth image curves at multiple scales. In *International Conference on Computer Vision*, pages 558–567, 1988.

[Lowe 91] D. Lowe. Fitting parametrized 3D models to images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(5):441–450, May 1991.

[Marr & Nishihara 78] D. Marr and H.K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London, Series B*, (200):269–294, 1978.

[Marr 82] D. Marr. *Vision.* Freeman, San Francisco, CA, 1982.

[Metaxas *et al.* 93] D. Metaxas, S.J. Dickinson, R.C., Munck-Fairwood, and L. Du. Integration of quantitative and qualitative techniques for deformable model fitting from orthographic, perspective and stereo projection. In *Fourth International Conference on Computer Vision*, pages 364–371, 1993.

[Metropolis *et al.* 53] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E.Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[Miglino *et al.* 96] O. Miglino, H. Hautop Lund, and S. Nolfi. Evolving mobile robots in simulated and real environments. *Artificial Life*, 1996. To appear.

[Minsky 75] M. Minsky. A framework for representing knowledge. In P.H. Winston, editor, *The Psychology of Computer Vision*. McGraw-Hill, 1975.

[Mohan & Nevatia 89] R. Mohan and R. Nevatia. Using perceptual organization to extract 3D structures. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11(11):1121–1139, 1989.

[Mohan & Nevatia 92] R. Mohan and R. Nevatia. Perceptual organization for scene segmentation and description. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 14(6):616–635, June 1992.

[Murase & Nayar 92] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appeareance. Technical Report CUCS-054-92, Columbia University (NY), 1992. To appear into the International Journal of Computer Vision.

[Nevatia & Binford 77] R. Nevatia and T.O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8:77–98, 1977.

[Pednault 89] E.P.D. Pednault. Some experiments in applying inductive inference principles to surface reconstruction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1603–1609, Detroit, MI, August 1989.

[Pentland & Sclaroff 91] A.P. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modelling and recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(7):715–729, July 1991.

[Pentland 86] A.P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986.

[Pentland 90] A.P. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, (4):107–126, 1990.

[Petitjean *et al.* 92] S. Petitjean, J. Ponce, and D.J. Kriegman. Computing exact aspect graphs of curved objects: Algebraic surfaces. *International Journal of Computer Vision*, 9(3):231–255, 1992.

[Pilu & Fisher 95] M. Pilu and R.B. Fisher. Equal-distance sampling of superellipse models. In *Proceedings of the British Machine Vision Conference*, volume 1, pages 257–266, Birmingham, September 1995.

[Pilu & Fisher 96a] M. Pilu and R.B. Fisher. Model-driven grouping and recognition of generic object parts from single images. In *4th International Symposium on Intelligent Robotic Systems*, pages 147–154, Lisboa, Portugal, July 1996.

[Pilu & Fisher 96b] M. Pilu and R.B. Fisher. Part segmentation from 2D edge images by the MDL criterion. In *Proceedings of the British Machine Vision Conference*, pages 83–92, Edinburgh, September 1996.

[Pilu & Fisher 96c] M. Pilu and R.B. Fisher. Part segmentation from 2D edge images by the MDL criterion. *Image and Vision Computing*, July 1996. Submitted.

[Pilu & Fisher 96d] M. Pilu and R.B. Fisher. Recognition of geons by parametrically deformable contour models. In R. Cipolla and B. Buxton, editors, *Fourth European Conference on Computer Vision*, volume I of *Lecture Notes in Computer Science*, pages 71–82, Berlin, April 1996. Springer-Verlag.

[Pilu & Fisher 96e] M. Pilu and R.B. Fisher. Recovery of generic parts by parametric deformable aspects. Technical Report 801, Department of Artificial Intelligence, University of Edinburgh, May 1996.

[Pilu & Mainardi 90] M. Pilu and F. Mainardi. *Analisi dell'incertezza nella localizzazione di oggetti a partire da immagini singole.* Tesi di Laurea, Dipartimento di Elettronica, Politecnico di Milano, Milano - Italy, December 1990. In Italian.

[Pilu *et al.* 96a] M. Pilu, A.W. Fitzgibbon, and R.B. Fisher. Ellipse-specific least-square fitting. In *IEEE International Conference on Image Processing*, Lausanne, Switzerland, September 1996.

[Pilu *et al.* 96b] M. Pilu, A.W. Fitzgibbon, and R.B. Fisher. Training PDM on models: The case of deformable superellipses. In *Proceedings of the British Machine Vision Conference*, pages 373–382, Edinburgh, September 1996.

[Pizer *et al.* 94] S.M Pizer, C.A. Burbeck, J.M. Coggins, D.S. Fritsch, and B.S. Morse. Object shape before boundary shape: Scale space medial axes. *Journal of Mathematical Imaging and Vision*, 4(3):303–313, 1994.

[Ponce *et al.* 92] J. Ponce, S. Petitjean, and K. Kriegman. Computing exact aspect graphs of curved objects: Algebraic surfaces. *Proceedings of the European Conference on Computer Vision*, pages 599–614, 1992.

[Porrill 90] J. Porrill. Fitting ellipses and predicting confidence envelopes using a bias corrected Kalman filter. *Image and Vision Computing*, 8(1):37–41, February 1990.

[Raja & Jain 92a] N.S. Raja and A.K. Jain. Obtaining generic parts from range data using a multi-view representation. In *Appl. Artif. Intell. X: Machine Vision and Robotics, Proc. SPIE 1708*, pages 602–613, Orlando, FL, April 1992.

[Raja & Jain 92b] N.S. Raja and A.K. Jain. Recognizing geons from superquadrics fitted to range data. *Image and Vision Computing*, 12(3):179–189, April 1992.

[Ramer 72] U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1:244–256, 1972.

[Rao & Nevatia 89] K. Rao and R. Nevatia. Descriptions of complex objects from incomplete and imperfect data. In *DARPA Image Understanding Workshop*, pages 399–414, San Mateo CA, 1989. Morgan Kaufmann.

[Rao 84] S. S. Rao. *Optimization: Theory and Applications.* Wiley Estern, 1984. 2nd edition.

[Rissnanen 83] J. Rissnanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 2:416–431, 1983.

[Roberts 65] L.G. Roberts. Machine perception of three-dimensional solids. In J.P. Tippett, editor, *Optical and Electro-Optical Information Processing.* MIT Press, Cambridge, MA, 1965.

[Rock 83] I. Rock. *The Logic of Perception.* MIT Press, 1983.

[Rom & Medioni 93] H. Rom and G. Medioni. Hierarchical decomposition and axial shape description. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(10):973–981, 1993.

[Rosch 73a] E. Rosch. Natural categories. *Cognitive Psychology*, (4):328–350, August 1973.

[Rosch 73b] E. Rosch. On the internal structure of perceptual and semantic categories. In T.E. Moore, editor, *Cognitive Development and the Acquisition of Language*. Academic Press, New York, 1973.

[Rosch 78] E. Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Erlbaum, Hillsdale, NJ, 1978.

[Rosin & West 90] P.L. Rosin and G.A. West. Segmenting curves into lines and arcs. In *Proceedings of the Third International Conference on Computer Vision*, pages 74–78, Osaka, Japan, December 1990.

[Rosin & West 95] P.L. Rosin and G.A. West. Non-parametric segmentation of curves into various representations. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(12):1140–1153, December 1995.

[Rosin 93a] P. L. Rosin. Elipse fitting by accumulating five-point fits. *Pattern Recognition Letters*, (14):661–699, August 1993.

[Rosin 93b] P.L. Rosin. A note on the least squares fitting of ellipses. *Pattern Recognition Letters*, (14):799–808, October 1993.

[Saint-Marc & Medioni 88] P. Saint-Marc and G. Medioni. Adaptive smoothing for feature extraction. In *Proceedings of DARPA Image Understanding Workshop*, pages 1100–1113, Cambridge(MA), 1988.

[Sakar & Boyer 93] S. Sakar and K.L. Boyer. Integration, inference, and management of spatial information using bayesian networks: Perceptual organization. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15(3):256–274, 1993.

[Sampson 82] P.D. Sampson. Fitting conic sections to very scattered data: an iterative refinement of the Bookstein algorithm. *Computer Graphics and Image Processing*, (18):97–108, 1982.

[Sclaroff & Pentland 95] S. Sclaroff and A. Pentland. Modal matching for corrispondence and recogntion. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(6):545–561, 1995.

[Scott & Longuet-Higgins 91] G.L. Scott and H.C. Longuet-Higgins. An algorithm for associating the features of two patterns. In *Proc. Royal Society London*, volume B244, pages 21–26, 1991.

[Sederberg *et al.* 84] T. W. Sederberg, D. C. Anderson, and R. N. Goldman. Implicit representation of parametric curves and surfaces. *Computer Vision, Graphics and Image Processing*, 28:72–84, 1984.

[Sha'ashua & Ullman 88] A. Sha'ashua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *International Conference on Computer Vision*, pages 321–327, 1988.

[Shapiro & Brady 92] L. S. Shapiro and J.M. Brady. Feature-based corresondence: an eigenverctor approach. *Image and Vision Computing*, pages 283–288, June 1992.

[Shapiro & Haralick 79] L. Shapiro and R. Haralick. Decomposition of two-dimensional shapes by graph theoretic clustering. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1(1):10–20, 1979.

[Siddiqi & Kimia 95] K. Siddiqi and B.B. Kimia. Parts of visual form: computational aspects. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(3):239–251, March 1995.

[Solina & Bajcsy 90] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case of superquadrics with global deformations. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(2):131–147, February 1990.

[Staib & Duncan 92] L.H. Staib and J.S. Duncan. Boundary finding with parametrically deformable models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, November 1992.

[Stein & Medioni 92] F. Stein and G. Medioni. Recognition of 3D objects from 2D groupings. In *Proceeding of the DARPA Image Understanding Workshop*, pages 26–29, 1992.

[Stewman & Bowyer 90] J. Stewman and K.W. Bowyer. Direct construction of perpective projection aspect graphs for planar-face convex objects. *Computer Vision, Graphics and Image Processing*, 51:20–37, 1990.

[Subirana-Vilanova & Richards 91] J. Subirana-Vilanova and W.A. Richards. Perceptual organization, figure-ground, attention and saliency. MIT AI Memo 1218, MIT, 1991.

[Subirana-Vilanova 93] J.B. Subirana-Vilanova. Mid-level vision and recognition of non-rigid objects. AI-TR 1442, MIT, January 1993.

[Tarr & Bulthoff 95] M.J. Tarr and H.H. Bulthoff. Is human object recognition better described by geon structural descriptions of by multiple views? - Comment Biederman and Gerhardstein (1993). *Journal of Experimental Psychology - Human Perception and Performance*, 21(6):1494–1505, 1995.

[Taubin 91] G. Taubin. Estimation of planar curves, surfaces and non-planar space curves defined by implicit equations, with applications to edge and range image segmentation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(11):1115–1138, November 1991.

[Teh & Chin 92] C. Teh and R.T. Chin. On the detection of dominant points on digital curves. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11(8):859–872, 1992.

[Ullman 79] S. Ullman. *The Interpretation of Visual Motion.* MIT Press, Cambridge, MA, 1979.

[Waltz 75] D. Waltz. Generating semantic descriptions from drawing of scenes with shadows. In P.H. Winston, editor, *Psychology of Computer Vision*, pages 19–92. McGraw-Hill, New York, 1975.

[Wertheimer 23] M. Wertheimer. Laws of organization in perceptual form. In W.D. Ellis, editor, *A Source Book of Gestalt Psychology.* Harcourt Brace, New York, 1923.

[Wilkinson 65] J. H. Wilkinson. *The algebraic eigenvalue problem.* Claredon Press, Oxford, England, 1965.

[Williams & Jacobs 95] L. Williams and D.W. Jacobs. Stochastic completion fields: A neural model of illusory contour shape and saliency. In *International Conference on Computer Vision*, pages 408–415, Cambridge, MA, 1995.

[Witkin & Tenenbaum 85] A.P. Witkin and J.M. Tenenbaum. On perceptual organization. In A. Pentland, editor, *From Pixels to Predicates.* Ablex, Norwood, NJ, 1985. Original work published in 1981.

[Wittgenstein 53] L. Wittgenstein. *Philosophical Investigation.* Macmillan, New York, 1953.

[Wren 96] D.O. Wren. *Planning Dextrous Grasps from Range Data using Preshaping and Digit Trajectories.* Unpublished PhD Thesis, Department of Artificial Intelligence, University of Edinburgh, May 1996.

[Wu & Levine 94] K. Wu and M.D. Levine. Recovering of parametric geons from multiview range data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 159–166, Seattle, WA, 1994.

[Yuille *et al.* 92] A. Yuille, P.W. Hallinan, and D.S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8:99–111, 1992.

[Zerroug & Nevatia 94] M. Zerroug and R. Nevatia. Three-dimensional part-based description from a real intensity image. In *ARPA Image Understanding Workshop*, pages 1367–1374, Monterrey, CA, November 1994.

[Zhu & Yuille 95] S. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, energy/Bayes/MDL for multi-band image segmentation. In *International Conference on Computer Vision*, pages 416–423, 1995.